N° d'ordre

THESE

présentée

A L'UNIVERSITE PARIS VII

pour obtenir le titre de

DOCTEUR INGENIEUR

SPECIALITE ELECTRONIQUE (TRAITEMENT DE L'INFORMATION)

par

Georges CARAYANNIS

ANALYSE DE LA PAROLE PAR IDENTIFICATION

RECURRENTE D'UN MODELE DU SYSTEME DE PHONATION

Soutenue le 30 Novembre 1973 devant la Commission d'examen:

MM. A. BLAQUIERE *Président*

P. DEBRAINE

H. BESTOUGEFF *Examinateurs*

C. GUEGUEN

E R R A T A

Page 59 : Formule (3.19) écrire :

$$\text{Log } (|E|(\omega)|^2) + \text{Log } (|H|(\omega)|^2)$$

Page 85 : 3 ème ligne écrire :

En effet, nous savons que (4 ÷ 5) formants ...

Page 98 : Formule 8 ème ligne écrire :

$$\frac{\delta\ \varepsilon^2}{\delta\ a} = 2\ (-S^T \underline{s} + S^T S\ \underline{a}) = 0$$

Page 102 : Formule (6.15) écrire :

$$R\underline{x} = \underline{v}$$

Page 104 : Formules (6.16) écrire :

$$p\ (s/r) = \frac{C}{(2\ \pi)^{N/2}\ |V|^{1/2}}\ \exp\ (-\frac{1}{2}\ \underline{S}^T\ V^{-1}\ \underline{s})$$

Page 118 : Formule 3 ème ligne écrire :

$$r_n = \sum_{j=0}^{N-1-n} \varepsilon_j \varepsilon_{j+n}$$

Page 130 : Formules (7.29) écrire :

$$\delta_n = \underline{\hat{a}}_n^T\ \underline{\hat{a}}_n - \underline{\hat{a}}_{n-1}^T\ \underline{\hat{a}}_{n-1}$$

$$\gamma_n = (\underline{\hat{a}}_n - \underline{\hat{a}}_{n-1})^T\ (\underline{\hat{a}}_n - \underline{\hat{a}}_{n-1})$$

THESE

présentée

A L'UNIVERSITE PARIS VII

pour obtenir le titre de

DOCTEUR INGENIEUR

SPECIALITE : ELECTRONIQUE (TRAITEMENT DE L'INFORMATION)

par

Georges **CARAYANNIS**

ANALYSE DE LA PAROLE PAR IDENTIFICATION

RECURRENTE D'UN MODELE DU SYSTEME DE PHONATION

Soutenue le 30 Novembre 1973 devant la Commission d'examen:

MM. A. BLAQUIERE *Président*

P. DEBRAINE

H. BESTOUGEFF *Examinateurs*

C. GUEGUEN

UNIV.
LEIDEN
BIBL.

A la mémoire de ma mère

Digitized by the Internet Archive
in 2026

https://archive.org/details/IA41548022_0021

La présente thèse a été réalisée au LABORATOIRE D'AUTOMATIQUE de l'E. N. S. T. que dirige Monsieur le Professeur C. GUEGUEN.

Je lui exprime ma profonde reconnaissance pour la sollicitude constante et les précieux conseils qu'il m'a prodigués tout au long de cette étude.

Je rends hommage à Monsieur le Professeur A. BLAQUIERE, Professeur d'AUTOMATIQUE THEORIQUE à l'Université PARIS VII qui a bien voulu présider le Jury.

Je témoigne ma profonde gratitude à Monsieur le Professeur P. DEBRAINE, Professeur à l'I. N. S. T. N., Adjoint au Directeur du Centre d'Etudes Nucléaires de SACLAY, pour le vif intérêt qu'il m'a manifesté et pour les précieuses remarques qu'il m'a apportées lors de nos rencontres.

Toute ma reconnaissance va à Mademoiselle BESTOUGEFF, Maître de Conférences à l'Université PARIS VII pour l'honneur qu'elle me fait en participant à mon Jury.

Je tiens à remercier, plus particulièrement, Monsieur MAISSIS, Ingénieur à l'E. N. S. T. pour les discussions constructives et son aide précieuse qu'il m'a manifestée tout au long de mon étude.

Enfin, je tiens à remercier tous ceux qui m'ont permis de mener à bien ce travail. Outre, l'environnement que j'ai trouvé au Laboratoire d'Automatique, je pense particulièrement, à Monsieur BACH, Ingénieur au Centre de Calcul de l'E. N. S. T. pour son assistance "système".

INTRODUCTION p 1

1. PRODUCTION DE LA PAROLE (PHONATION) p 7

1.1 Fonctionnement des organes de phonation p 7
1.2 Les sons de la parole p 9
1.3 Classification des sons p 10

1.3.1 Les voyelles p 11
1.3.2 Les consonnes p 11

1.4 Modèle à paramètres localisés de l'appareil vocal p 15
1.5 Modèle à paramètres répartis de l'appareil vocal p 18
1.6 Caractérisation mathématique des voyelles p 25

2. SIGNAL VOCAL - DIFFERENTS SUPPORTS D'INFORMATION p 28

3. ANALYSE DE LA PAROLE - METHODES CLASSIQUES D'ANALYSE p 44

3.1 Définitions p 44
3.2 Méthodes Analogiques p 46
3.3 Méthodes Numériques p 49

3.3.1 Généralités p 49
3.3.2 Analyse en Fréquence p 49

3.3.2.1 Définitions , Spectre à court terme, Spectre évolutif p 49
3.3.2.2 Recherche de l'information phonémique dans le spectre , cepstre , enveloppe p 57
3.3.2.3 Applications p 66

A. Détection automatique des formants p 66
A1. Chirp - z - tranform p 67

A2. Détection des formants avec un modèle du spectre p 71
B. Synthèse de la parole p 73

3.3.3 Méthodes temporelles p 80

3.3.3.1 Méthode des passages à zéro p 80

3.3.4 Conclusion p 84

4. RECHERCHE DES NOUVELLES METHODES p 86

4.1 Parametrisation de la parole en utilisant un modèle de la membrane basilaire p 86
4.2 Application à la reconnaissance de la parole p 89

5. ANALYSE DE LA PAROLE EN UTILISANT UN MODELE - IDENTIFICATION DU SYSTEME DE PHONATION p 90

5.1 Définitions p 90
5.2 Modèlisation p 91

6. METHODES GLOBALES D'IDENTIFICATION p 97

6.1 Méthode dite de la prédiction linéaire p 97
6.2 Méthode de maximum de vraisemblance p 104
6.3 Filtrage inverse p 115
6.4 Implantation asynchrone des méthodes d'identification. Mesure du fondamental. p 116

7. ANALYSE DE LA PAROLE PAR IDENTIFICATION RECURRENTE p 119

7.1 Généralités p 119
7.2 Le filtre de Kalman Discret p 120
7.3 Application à l'analyse du signal vocal p 123
7.4 Mise en oeuvre de l'algorithme p 126

7.5 Applications pratiques du modèle p 130
7.6 Discussion p 141
7.7 Réalisations - programmes p 144

8. SYNTHESE DE LA PAROLE p 146

8.1 Généralités p 146
8.2 Modèles de synthèse p 149
8.3 Synthètiseur de la parole attaché à un analyseur " Kalman " p 152

9. TRANSITIONS PHONEMIQUES p 154

9.1 Le rôle des transitions phonèmiques dans la structure d'une phrase p 154
9.2 Modèlisation des transitions p 155
9.3 Application à la segmentation de la parole p 157
9.4 Extensions - Conclusions . p 161

CONCLUSION p 162

BIBLIOGRAPHIE p 165

ANNEXES p I

ANNEXE I : Analyse de differents phonèmes avec la méthode de deconvolution du signal vocal p I

ANNEXE II : Analyse de la parole avc le filtre de Kalman p IX

ANNEXE III : Programmes p XXIII

ANNEXE IV : Comparaison des méthodes p ILI

I N T R O D U C T I O N

La parole est le moyen de communication le plus naturel de l'homme. L'utilisation de ce support pour la communication avec la machine semble donc très souhaitable. Ce dialogue peut se faire dans deux sens (homme → machine et machine → homme). Ainsi, deux unités différentes sont nécessaires pour assurer l'interface : une unité de reconnaissance de la parole et une unité de synthèse vocale.

Toutes les applications se trouvent conditionnées par une étape importante qui est l'analyse de la parole (irréversible, rarement remise en question). Il nous a donc paru important, bien qu'un certain nombre de techniques existent déjà, de rechercher et d'évaluer la puissance de nouvelles techniques d'analyse.

Le signal vocal est un signal très riche en diverses informations. Une représentation numérique directe de ce signal après conversion analogique-digitale aboutirait à des taux d'information très élevés, prohibitifs pour les différentes applications, où tout traitement se heurterait au grand volume des données brutes. L'analyse consiste à extraire du signal acoustique un nombre réduit de paramètres pertinents décrivant certains aspects de ce signal.

Toute analyse efficace se fait donc au prix d'une réduction des données et doit être adaptée à l'extraction du type d'information recherchée. Ainsi, les paramètres ne sont "pertinents" ou redondants" qu'en fonction d'une application postérieure. Souvent, c'est l'information "phonémique" (ou sémantique) qui est recherchée. Dans ce cas, une invariabilité par rapport au locuteur (par exemple) est souhaitable.

L'analyse de la parole est donc un problème délicat et non encore complètement résolu car aucune méthode actuelle n'assure ni une séparation complète entre les différents types d'information, ni une extraction exhaustive de l'information choisie.

L'importance de l'analyse à travers ses diverses applications est manifeste :

Un premier aspect du dialogue homme-machine est la synthèse de la parole. Une voix "standard" peut en général être acceptée. La qualité de la voix engendrée dépend de l'unité informationnelle que l'on utilise à la synthèse (phonème, diphonème, mot, etc ...) ainsi que de la qualité de représentation du signal après analyse. L'intérêt d'utilisation d'unités réduites dont la juxtaposition conduit au signal synthétique, est lié au problème de mémorisation des données (par exemple, le nombre de diphonèmes est limité par rapport au nombre total des mots pour une langue). La juxtaposition devient d'autant plus difficile que l'unité informationnelle est réduite (par exemple : phonème). Des règles de passage de nature certainement complexe doivent être recherchées dans ce cas.

Le deuxième aspect du dialogue est fondée sur la nécessité, pour la machine, de pouvoir reconnaître ce qui a été dit par l'homme. On peut distinguer trois niveaux de communication :

a - Simple commande d'automate : un dictionnaire très limité à quelques mots-clés est nécessaire dans ce cas. Le problème est relativement simple. L'analyse du signal qui doit être réalisée peut être assez restreinte.

b - Communication de haut niveau dans le sens homme-machine :
Le caractère conversationnel n'est pas essentiel ici. On désire demander à la machine certains travaux plutôt simples : par exemple, modifier le fichier contenant l'état de stocks d'un magasin au fur et à mesure que les produits se vendent, établir les opérations de caisse correspondantes, etc ... Dans ce cas, le vocabulaire est plus étendu tandis que l'environnement syntaxique peut demeurer limité.

c - Communication dans les deux sens : homme - machine et machine - homme
A ce niveau, la possibilité de synthèse de la parole est aussi essentielle. La machine doit répondre aux ordres donnés. Elle doit pouvoir prendre des décisions et pouvoir les communiquer directement à l'homme d'une façon claire. Une telle machine doit comporter deux unités : une unité de reconnaissance et une unité de réponse vocale.

Des applications importantes peuvent être envisagées à ce niveau, comme la programmation directe des ordinateurs, la communication avec des robots, etc ... Ici, on doit être très éxigeants en ce qui concerne la qualité de reconnaissance. Cette qualité est liée à la qualité des paramètres, donc avec la qualité de l'algorithme de paramètrisation. Ceci parce que le dictionnaire étant assez étendu (par exemple, langage de programmation évolué, conversation courante) une analyse plus fine du signal vocal doit précéder la reconnaissance.

La reconnaissance de la parole est considérée aujourd'hui comme un problème très difficile à résoudre et ceci parce qu'il y a beaucoup de différences entre les voix des diverses personnes. Une normalisation alors des voix par rapport au locuteur serait souhaitable avant reconnaissance. Ceci pour ne pas construire des machines qui reconnaissent uniquement "la voix de leur maître" comme on fait actuellement.

Nous avons évoqué le caractère individuel de la parole. En effet, les différences entre les voix des différentes personnes sont si importantes qu'à chaqune, un et un seul individu peut être associé. Un autre domaine de recherche se découvre alors, celui de la vérification et de l'identification des locuteurs avec leurs applications respectives. La question qui se pose tout de suite est la suivante : EST-CE QUE L'INFORMATION SUR LA PERSONNE SE RETROUVE "A COURT TERME" sur LE SIGNAL VOCAL ? C'est-à-dire est-ce qu'à un segment de signal de 10 ms, par exemple, on peut associer le concept de locuteur individualisé aussi bien que le concept de phonème ? Les recherches qui ont été entreprises ont abouti à la conclusion que le contenu informationnel de certains phonèmes, comme par exemple, les nasales, était très sensible au locuteur. Pourtant, ce résultat ne peut pas être généralisé. Ainsi, nous savons qu'un effort maximum est fait par chacun pour s'adapter à la prononciation de chaque phonème d'une façon"standard"dictée par le mécanisme auditif. Ceci n'est pas aussi facile pour tous les phonèmes étant données certaines particularités à la construction de l'appareil de phonation qui font que certains gardent plus de traits individuels que d'autres. Il en est de même pour quelques défaillances du mécanisme cybernétique qui contrôle le processus

de la parole. La recherche de cette information à court terme liée avec la personne qui a parlé, parait souhaitable pour normaliser les paramètres de reconnaissance. Pourtant, elle est difficilement exploitable pour en bâtir des systèmes d'identification automatique des locuteurs. Ainsi, c'est plutôt l'évolution de certains paramètres tout le long d'une phrase-code qui est utilisée. De cette façon, un très grand volume de données de base est nécessaire. Il y a ici une différence essentielle avec les systèmes de reconnaissance des phonèmes où l'analyse est plus rudimentaire.

La seconde différence est liée avec le nombre de locuteurs qui participeront à l'apprentissage. Ce nombre peut être fixé assez arbitrairement pour chaque application à l'encontre du nombre de phonèmes qui est fixé une fois pour toutes. On verra, au cours de cet exposé (CH 2), que l'évolution de certains paramètres simples peut être utilisée dans le cadre de l'identification des locuteurs.

Une autre application, où l'analyse de la parole joue un rôle essentiel, est la transmission numérique. Celle-ci revêt deux aspects importants :

a) la réduction de taux d'information à transmette ;

b) la préservation du secret des communications.

C'est la représentation globale du signal qui est recherchée ici. L'ensemble des opérations,analyse et synthèse,doit pouvoir se faire en temps réel.

Dans ce rapport de travail, nous exposons une nouvelle technique d'analyse de la parole. Cette technique est basée sur l'utilisation du filtre de KALMAN discret. Le filtre de KALMAN est introduit dans un processus d'identification de la fonction de transfert du canal vocal par la méthode du modèle. Un modèle général d'une fonction de transfert en z est adopté. Cette méthode s'est avérée très efficace pour analyser le signal vocal au point de vue théorique ainsi qu'au point de vue pratique. L'aspect pratique est très important parce que l'automatisation complète de l'algorithme est possible. Il nous a paru intéressant de pouvoir comparer cette méthode avec d'autres déjà existantes. Pour faire ceci, nous avons programmé deux autres méthodes. La première est celle dite de

prédiction linéaire basée sur une approche des moindres carrés. L'autre est la méthode du cepstre (approche fréquentielle). Une grande partie de ce rapport donne les détails techniques et théoriques de ces méthodes, ceci pour pouvoir se placer dans un contexte plus général.

Au CH 1 du rapport, nous exposons brièvement ce qui concerne la physiologie des organes vocaux et la production de la parole. Ceci a été inévitable, au moins pour pouvoir donner quelques définitions.

Au CH 2, nous essayons de décrire le signal vocal ; cette description vise à séparer déjà le signal en plusieurs types de supports informationnels, en vue de faciliter l'étude qui suit.

Au CH 3, nous exposons les méthodes classiques d'analyse. Nous examinons en détail la méthode du cepstre.

Au CH 4, nous résumons une méthode de paramètrisation utilisée au Laboratoire d'Automatique de l'E. N. S. T. en vue d'une reconnaissance de la parole. L'approche adoptée est basée sur un modèle du système auditif.

Au CH 5, nous discutons l'aspect "Identification du système de Phonation" et au CH 6, nous exposons certaines techniques d'identifications appliquées avec succès au traitement du signal vocal.

Au CH 7, nous exposons la technique d'analyse que nous avons développée au Laboratoire d'Automatique de l'E. N. S. T. basée sur l'utilisation du filtre de KALMAN pour identifier le système de phonation. Une estimation des paramètres du modèle utilisé peut être fournie à chaque instant. Ceci permet la détection simultanée du fondamental grâce à un critère intrinsèque au système. En outre, une comparaison de la méthode avec les autres méthodes est effectuée par la suite afin de rechercher ses avantages et ses inconvénients.

Au CH 8, nous abordons le problème de la synthèse de la parole en général. Nous voyons la synthèse comme application directe de l'analyse.

Nous exposons en outre comment fonctionne notre synthétiseur qui est simulé sur CII 10070. Un modèle abstrait sous forme d'une relation de récurrence entre les échantillons successifs du signal est adopté.

Au CH 9, nous examinons le rôle des transitions phonémiques sur la parole et nous proposons un détecteur automatique des transitions.

Beaucoup de travail reste encore à faire pour arriver à maîtriser le signal de parole. Le développement des techniques numériques pour l'analyse du signal vocal est très prometteur. Dans leur cadre, la recherche de l'information semble systématisée. Un grand nombre d'opérations numériques et logiques peuvent être effectuées facilement en ce cas et des décisions prises durant l'analyse du signal. Ainsi, le mécanisme cybernétique de production se découvre au fur et à mesure.

Nous espèrons que ce travail a pu contribuer à quelques progrès dans ce domaine.

1. PHONATION.

1.1 Fonctionnement des organes de phonation.

Au niveau acoustique, le signal vocal est constitué par de rapides fluctuations de la pression de l'air (ondes sonores) engendrées par l'appareil vocal . Celui-ci se compose d'un réservoir d'énergie constitué par le système respiratoire et d'un ensemble de sources et de cavités de résonance dont les mouvements volontaires produisent et modulent la parole. Le processus complet de la phonation apparaît comme un système cybernétique asservi dans la mesure où la parole est développée, controlée et entretenue par l'intervention d'une boucle de retour : le système auditif.

La figure 1.2 présente un diagramme schématique des organes phonatoires. Le canal vocal apparaît comme un conduit de section variable qui s'étend des cordes vocales aux lèvres . Sa longueur est environ de 17 cm pour un adulte et sa section varie de zéro à 20 cm^2 . Les cavités nasales constituent un conduit additionnel qui intervient en dérivation sur le précédent pour la production des sons nasalisés . Ce canal s'étend de la luette aux narines . Sa longueur pour l'adulte atteint 12 cm et son volume environ 60 cm^3 . Le degré de couplage entre ces deux canaux est commandé par l'ouverture de la luette .

Le système respiratoire fournit l'énergie nécessaire à la production de la parole en forçant les vibrations de l'appareil vocal . Il est commode de considérer trois sources qui représentent les aspects principaux que revêt le mode d'excitation . Ces sources sont les suivantes:

i. Première source: Les vibrations des cordes vocales donnent naissance aux sons voisés, ces vibrations sont quasi-périodiques et sont constituées d'une succession de cycles. Chaque cycle se déroule comme suit : la pression de l'air sous-glottale augmentant suffisamment , force les cordes à se séparer , ces cordes étant jointives au repos ; un flux d'air traverse les cordes vocales et , selon la loi de Bernoulli, la pression diminue et une force de rappel ramène les cordes à leur état initial .

Ces oscillations sont du type forcé et leur période est essentiellement fonction de la masse et de la tension des cordes ainsi que de la pression sous glottale.Le flux d'air s'échappant de façon impulsionnelle

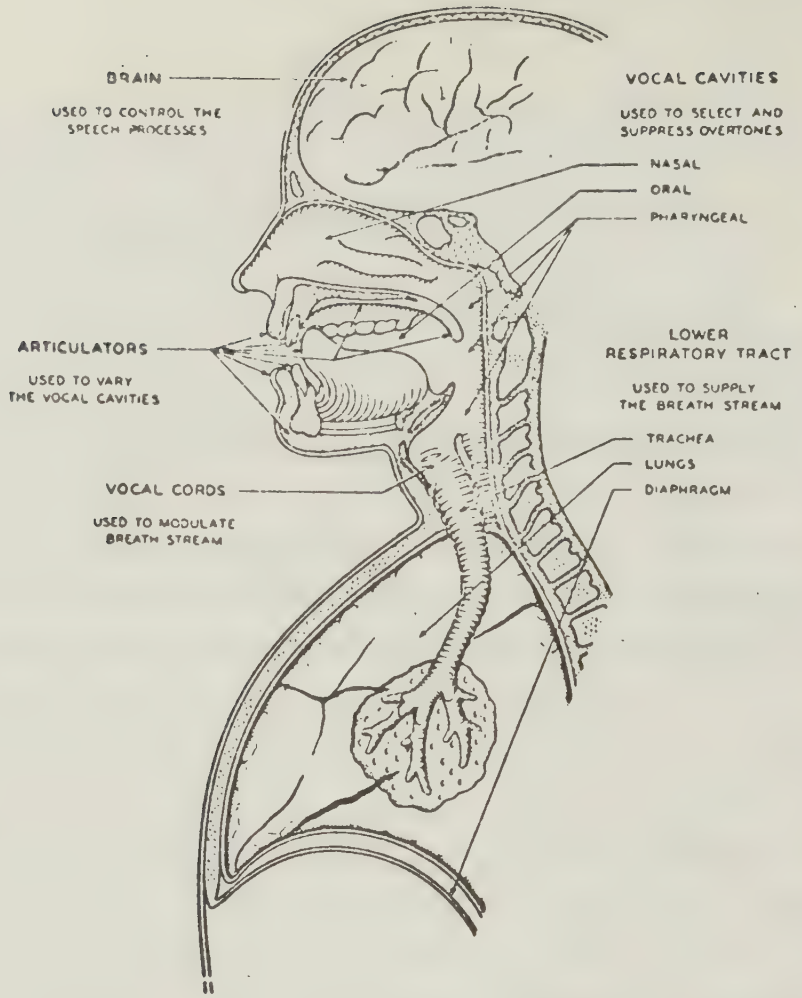


Figure 1.1. :

Facteurs dont dépend la phonation . Plusieurs organes sont impliqués à ce processus. (Figure tirée de /51/.)

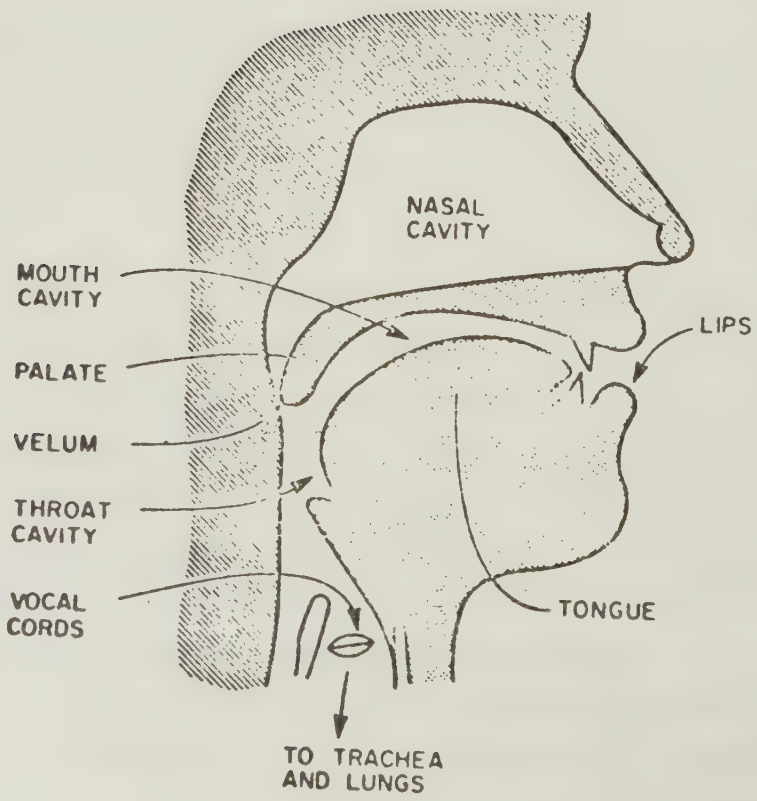


Figure 1.2. :

Diagramme schématique de l'appareil vocal. (tiré de /52/)

fait vibrer le système vocal à ses fréquences naturelles et il en résulte une distribution spectrale d'énergie . Dans ce spectre , on distingue des zones de résonance à forte concentration d'énergie qui constituent les "formants".

La propriété essentielle de cette source est sa périodicité (ou plutôt sa quasi-périodicité) exprimée par la période fondamentale $F_o = 1/ T_o$.

ii. Seconde source : Cette source est formée par un retrécissement du canal vocal en un de ses points. Il se crée en ce point un flux d'air turbulent qui engendre une excitation incoliérente du système vocal.

Puisqu'au point de sa création, le bruit est à spectre relativement uniforme sur une large bande , le spectre du son émis est déterminé par les cavités vocales qui se trouvent loin du point de rétrécissement .

iii. Troisième source : Cette troisième source est formée par une occlusion en un certain point du canal vocal : une pression étant maintenue par les poumons sur cette paroi , l'ouverture de cette dernière crée une excitation transitoire du canal vocal . D'une façon très schématique , l'excitation apériodique créée par cette source peut être considérée comme l'excitation du canal vocal à l'aide d'une fonction échelon ayant en conséquence un spectre qui décroit rapidement en fonction de la fréquence.

1.2 Les sons de la parole.

Au niveau sémantique, la langue parlée est un code continu, mais malgré la flexibilité de l'appareil vocal et malgré la nature continue de l'onde acoustique , l'homme est capable de discrétiser au niveau de la perception ce code en le segmentant en " phonèmes". Les "phonèmes" sont définis comme étant les sons simples dont l'enchaînement forme le langage : la représentation du langage se fait donc à l'aide des unités linguistiques de base que représentent les phonèmes . Ces phonèmes forment un code discret et le remplacement d'un unité linguistique par uneautre peut modifier le sens de l'expression ou lui ôter tout sens.

Le caractère plus ou moins prévisible du mot prononcé (contexte , grammaire...) permet souvent de se contenter d'une articulation approximative tout en assurant sa compréhension par l'auditeur ; dans beaucoup de cas,en effet , un seul mouvement vers l'état propre d'un

phonème peut être suffisant pour son articulation.. Il faut aussi remarquer que les relations entre les sons du langage et les mouvements correspondants des cavités vocales ne sont pas biunivoques , bien que les locuteurs normaux opèrent avec une grande similitude . Les variantes que peut présenter un même phonème chez différents locuteurs sont appelées "allophones " ; or,si on considère les phonèmes comme un code relié d'une façon unique aux articulations de l'appareil vocal, les allophones montrent qu'il y a une marge très large de choix lors de la détermination du code discret formé de phonèmes.

1.3 Classification des sons.

Les sons sont classés par les linguistes en fonction de la partie de l'appareil vocal qui les a émis et selon le mode d'articulation de tout l'appareil vocal. On opère donc deux distinctions:

- Les sons voisés (sonores) et les sons non-voisés (sourds) selon qu'ils mettent en jeu ou non la vibration des cordes vocales.

- Les sons continus et les sons non continus. Les sons continus peuvent être émis en l'absence de tout mouvement de l'appareil vocal (voyelle), les sons non continus par contre , mettent nécessairement en jeu un mouvement de l'appareil vocal, ils ont un caractère transitoire.

Au niveau de l'onde acoustique G. PETERSON /2/ classifie les sons en quatre classes , une classe acoustique étant définie comme l'ensemble des portions temporelles du signal qui ne font intervenir qu'une seule forme acoustique :

- Les sons quasi périodiques ou sons sonores dont les caractéristiques essentielles sont la fréquence fondamentale et les fréquences de résonance.

- Les sons quasi aléatoires produits par la seconde source d'excitation : le spectre est continu et les sons sont essentiellement caractérisés par

les domaines de résonance.

- les lacunes ou "gaps" : ces silences sont dus à l'occlusion de la troisième source.

- les impulsions après lacune dues à l'ouverture de la troisième source.

Les deux dernières classes sont caractérisées par la puissance instantanée du signal.

Si nous nous inspirons de la classification générale du langage écrit, nous pouvons opérer également un rangement des phonèmes, qui a pour intérêt de coïncider avec la classification générale utilisée en linguistique.

1.3.1. Les voyelles (voir tableau 1)

Pour une articulation normale, les cavités vocales restent dans une configuration relativement stable pour la plupart des sons continus et sonores qui constituent les voyelles. L'intervention des cavités nasales est négligeable et la diffusion se fait par la bouche.
En français il y a cependant des voyelles faisant intervenir les cavités nasales.

1.3.2. Les consonnes (voir tableau 1)

Les caractéristiques générales des consonnes sont les suivantes :

- emission non exclusivement par la bouche
- les cavités vocales ne demeurent pas dans une configuration stable
- le comportement dynamique de l'appareil vocal est décisif (régime transitoire)
- en plus d'une source spécifique, les cordes vocales peuvent intervenir.

TABLEAU 1

LISTE DES PHONEMES DU FRANÇAIS.

Phonème (1)	Mot clef	représentation machine	nature
a	plat	A } (2)	Voyelles
ɑ	mât	AA }	
i	il	I	
y	nu	Y	
ɔ	bol	O	
o	eau	OO	
ə	le	EE	
ɛ	lait	AI	
e	blé	E	
ø	peu	EU } (2)	
œ	heure	OE }	
u	ou	OU	
ã	an	AN	Voyelles nasales
ɔ̃	on	ON	
ɛ̃	lin	IN } (2)	
œ̃	brun	OEN }	
v	vie	V	F ricatives sonores
z	zéro	Z	
ʒ	je	ZZ	
f	feu	F	Fricatives sourdes
s	son	S	
ʃ	chat	SS	
b	bon	B	Occlusives sonores
d	dans	D	
g	gare	G	
p	pas	P	Occlusives sourdes
t	tas	T	
K	coût	K	

Phonème	Mot clef	Représentation machine	nature
m	ma	M	Consonnes nasales
n	nous	N	
ɲ	agneau	GN	
ŋ	camping	NG	
R	rue	R	Liquides
l	lent	L	
w	voir	W	Semi-voyelles
j	bailler	J	
ɥ	huit	HU	

(1) Symbolisme d'après l' IPA (International Phonetic Association)

(2) Phonèmes pratiquement indistinguables

On distingue :

a. Les consonnes fricatives.

- sonores ou sourdes selon qu'elles émanent de la première ou de la seconde source
- continues
- diffusées par la bouche exclusivement
- faible puissance acoustique

b. Les consonnes occlusives

- produites par la troisième source
- leur nature dépend du comportement dynamique des cavités
- sonores ou sourdes
- faible puissance et spectre très étalé

c. Les consonnes nasales

- transmission et émission essentiellement par les cavités nasales
- sonores
- continues

d. Les liquides et semi-voyelles

- très proches des voyelles
- sonores
- non intervention effective des cavités nasales

1.4. Modèle à paramètres localisés de l'appareil vocal.

La description quantitative de la production de la parole nécessite l'élaboration d'un modèle de l'appareil vocal. Etant donné la complexité des phénomènes mis en jeu, la mise en équation directe est impossible . On est donc réduit à approximer le comportement de l'appareil vocal par divers modèles dont la précision doit être en accord avec l'application envisagée (synthèse vocale , caractérisation du signal, mécanismes de fonctionnement du canal vocal ....). Dans de nombreux cas, un modèle grossier sera suffisant.

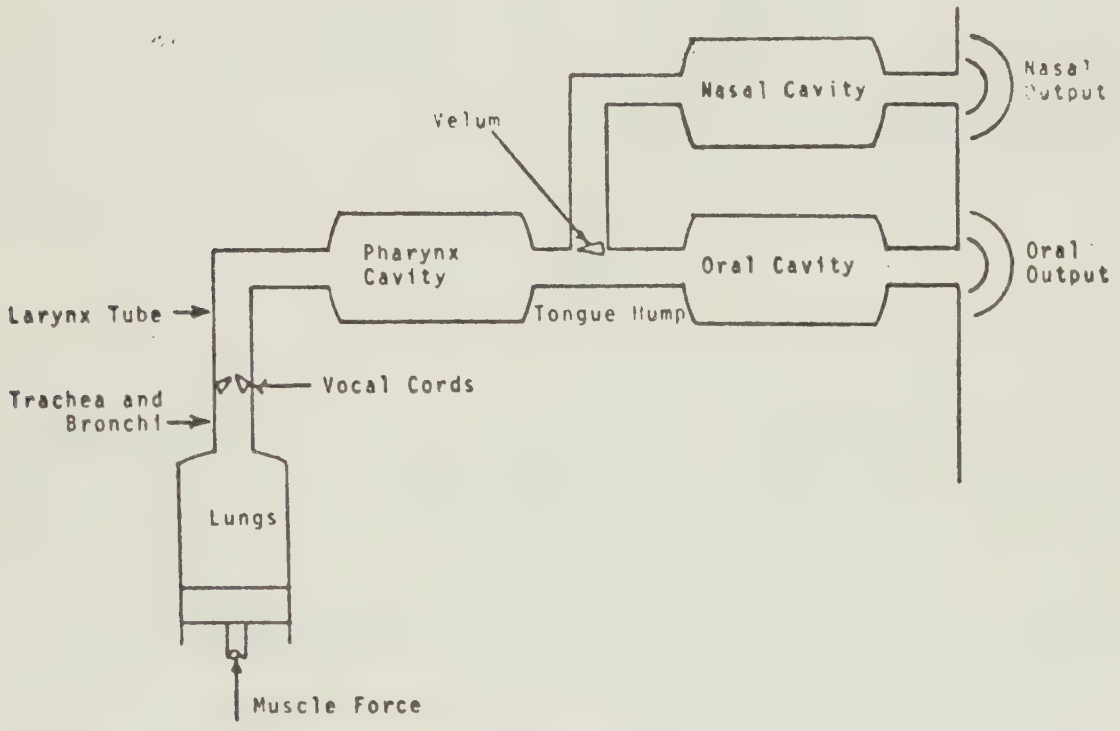


Figure 1.3. : Diagramme fonctionnel de l'appareil vocal. (tiré de /52/)

La figure 1.3. explicite le fonctionnement précedemment décrit . Utilisant l'analogie électrique , on peut considérer le signal vocal comme la sortie d'un filtre excité par diverses sources. Il en résulte un modèle électrique sous forme de quadripôle du type de la figure 1.4.

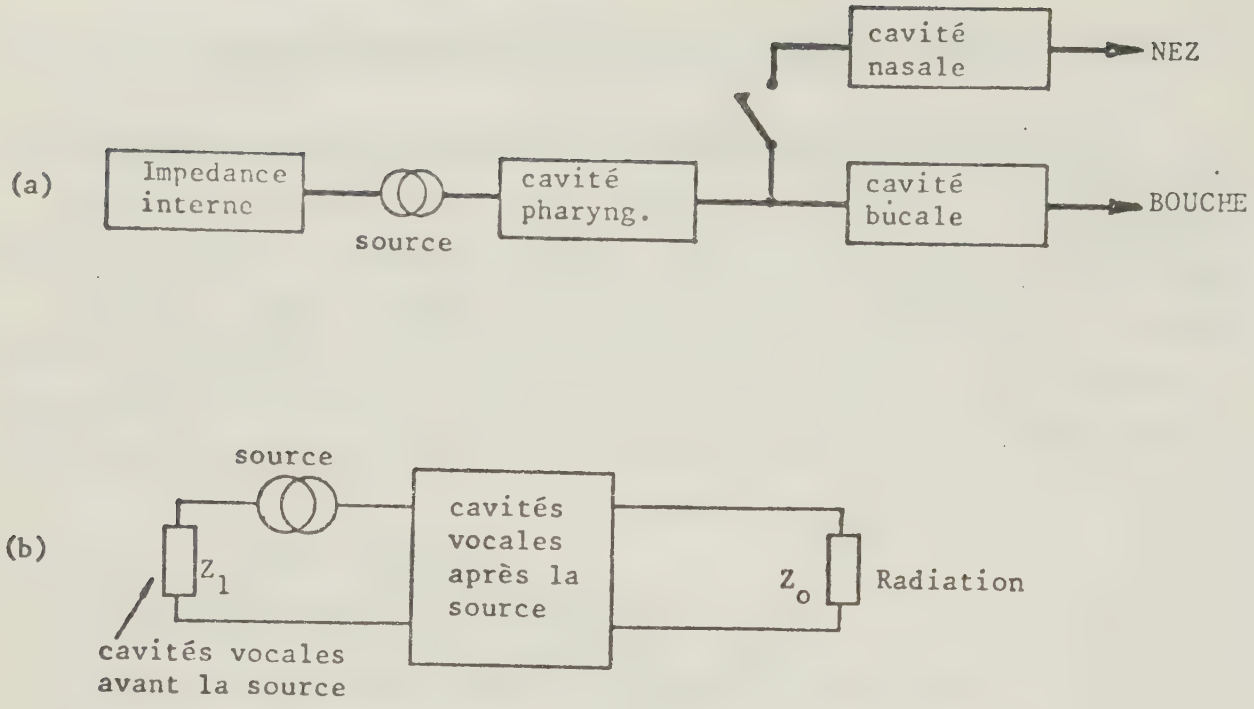


Figure 1.4. a - système de phonation sous forme de blocs
b - quadripôle équivalent.

Le système de phonation, ainsi considéré comme un quadripôle fermé par des impédances, est un filtre qui peut être caractérisé par sa fonction de transfert H(p)

Soient E(p) et S(p) les transformées des signaux de source et de sortie, on a :

$$S(p) = H(p)\, E(p) \qquad (1.1.)$$

ou encore si h(t) est la réponse impulsionnelle du filtre :

$$s(t) = h(t) * e(t) \qquad (1.2.)$$

La réponse harmonique du système se déduit de (1.1) selon la décomposition de la figure (1.5.)

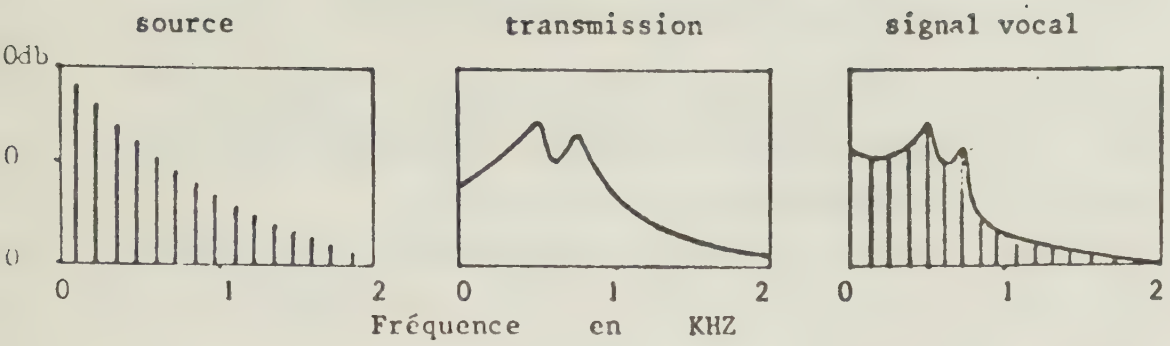


Figure 1.5 : Réponse spectrale de l'appareil vocal.

Le spectre du signal vocal fait donc intervenir le spectre de la source et la transmission du canal vocal. Les maxima de ce spectre sont dénommés FORMANTS . Ils caractérisent les propriétés du canal vocal et non de la source (ne coincident en général pas avec des harmoniques). On associe à chaque formant une bande passante à 3 db mesurant l'acuité du maximum.

La source d'excitation (i) est dans le cas des sons sonores (voisés) un signal quasi- périodique de type impulsionnel (ouverture brusque des cordes vocales). La fréquence correspondante est appelée FONDAMENTAL . Cette source est souvent approximée dans le modèle précédent par une série d'impulsions de dirac . Dans le cas de sons sourds, les sources de bruit (ii) et (iii) sont souvent approximées par un signal aléatoire blanc gaussien

A titre indicatif, on trouvera ci- après la marge de variation de la fréquence fondamentale ainsi que la valeur des 4 premiers formants pour des sons sonores non nasalisés prononcés par un locuteur masculin moyen /l/.

F_o = 60 - 240 Hz
F_1 = 150 - 850 Hz
F_2 = 500 - 2500 Hz
F_3 = 1500 - 3500 Hz
F_4 = 2500 - 4500 Hz

Le fondamental est , pour les locuteurs féminins, en général supérieur d'une octave, tandis que les fréquences des formants se révèlent seulement 17% plus élevés.

1.5 Modèle à paramètres répartis de l'appareil vocal

Un modèle plus réaliste est obtenu en considérant le canal vocal comme un système de propagation des ondes sonores. L'étude de ces propriétés de transmission, nécessite plusieurs étapes :

- découper le conduit en segments élémentaires suffisamment petits pour que leur section puisse être considérée comme constante. Cette opération est en général réalisée à partir de photos aux rayons X /1/.
- évaluer les propriétés de transmission de chaque segment.
- associer les divers segments. Les ordres de grandeur des longueurs d'onde permettent de supposer le canal linéaire et symétrique par rapport à sa ligne médiane.
- déterminer la fonction de transfert globale en tenant compte de la source et de l'impédance de radiation (bouche et nez).

Cette suite d'opérations est évidemment très complexe, d'autant plus que beaucoup de sons ont un caractère purement transitoire. Mais une étude de ce type même menée d'une façon très approximative, n'est pas dépourvue d'intérêt ; elle permet la mise en évidence des propriétés fondamentales du signal sonore.

A titre indicatif, nous allons étudier un cas très simplifié (/2/, /3/); le cas où le système vocal est assimilé à un tube d'une section constante A et d'une longueur l. Nous considérons que la transmission se fait par des ondes planes de telle façon que la pression et la vitesse volumique de l'onde ne dépendent que de la distance x (Figure 1.6).

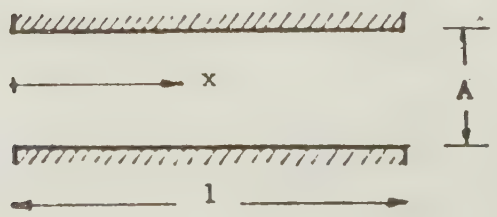


Figure 1.6.

Il s'agit d'une transmission avec des pertes d'énergie qui sont dues au frottement visqueux (pertes proportionnelles au carré de la vitesse) et à la chaleur dégagée (proportionnelle au carré de la pression). Le phénomène analogue sur le plan électrique , est celui d'une ligne de transmission où la tension représente la pression et le courant la vitesse.

Une ligne de transmission est caractérisée par (Figure 1.7.) :

- L'impédance par unité de longueur. $Z=R+j\omega L$
- L'admittance " " $Y=G+j\omega C$

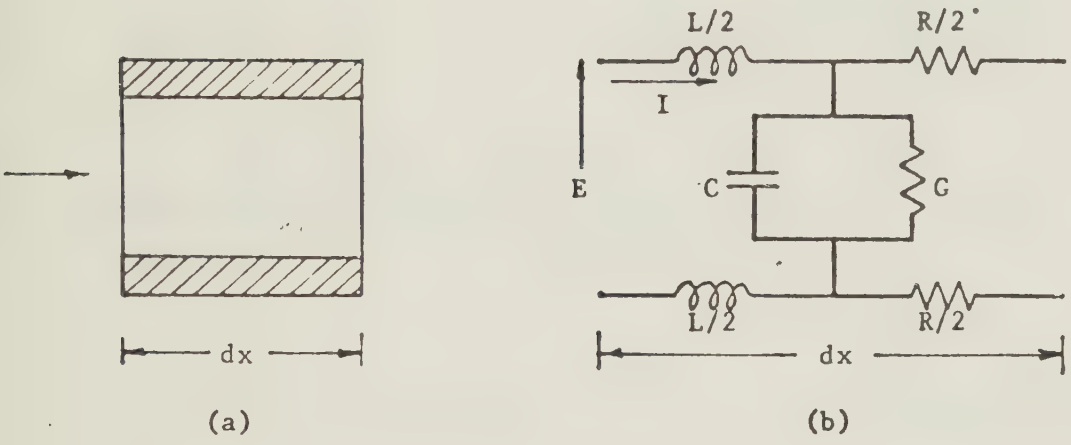


Figure 1.7. Caractéristiques du tube acoustique

(a) Représentation acoustique
(b) Ligne de transmission équivalente

Pour une excitation sinusoïdale du système (tension $Ee^{j\omega t}$, courant $Ie^{j\omega t}$) les équations différentielles de la ligne s'écrivent :

$$\frac{d^2E}{dx^2} - Z.Y.E = 0 \qquad E, I : E(x), I(x)$$

$$\frac{d^2I}{dx^2} - Z.Y.I = 0 \qquad (1.3.)$$

leurs solutions sont :

$$E = A_1 e^{\gamma x} + B_1 e^{-\gamma x}$$
$$I = A_2 e^{\gamma x} + B_2 e^{-\gamma x} \qquad (1.4.)$$

avec $\gamma = \sqrt{ZY} = \alpha + j\beta$ constante de propagation.

Les constantes des solutions(1.4) sont déterminées par les conditions aux limites de la ligne.Ainsi pour une longueur l, pour un couple tension/coura E_1, I_1, au départ et E_2, I_2 à l'arrivée les équations (1.4.)prennent la forme :

$$E_2 = E_1 ch(\gamma l) - I_1 Z_0 sh(\gamma l)$$
$$I_2 = I_1 ch(\gamma l) - E_1 Y_0 sh(\gamma l) \qquad (1.5.)$$

avec $Z_0 = Z/Y$, $Y_0 = Y/Z$ impédance et admittance caractéristiques de la ligne.

Les équations (1.5) écrivent :

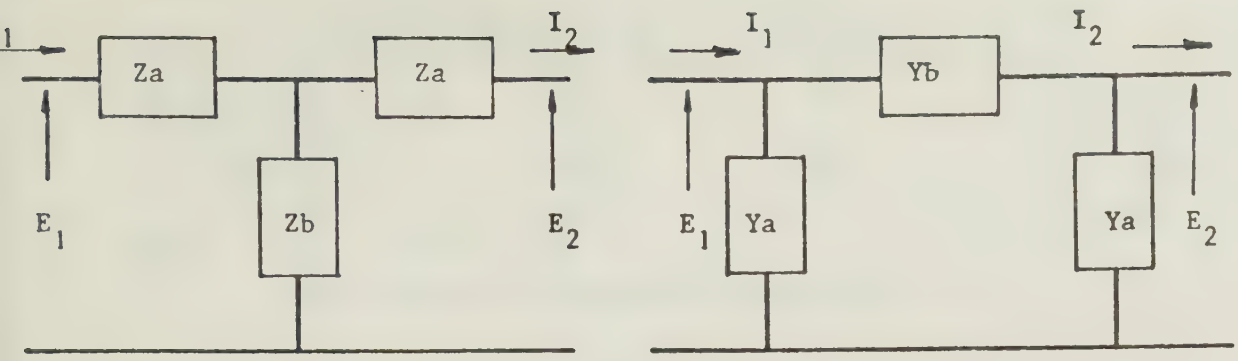
$$E_1 = I_1 \frac{Z_0}{sh(\gamma l)} - I_2 coth(\gamma l)$$
$$E_2 = I_1 Z_0 coth(\gamma l) - I_2 \frac{Z_0}{sh(\gamma l)} \qquad (1.6.)$$

Cette forme met en évidence les paramètres du quadripôle équivalent de la ligne.La figure 1.8. présente les caractéristiques de ce quadripôle en T et Π.

Rappelons que pour des pertes petites ($G << \omega C$, $R << \omega L$), la théorie des lignes déduit les formules suivantes:

$$\alpha \simeq \frac{R}{2}\sqrt{\frac{C}{L}} + \frac{G}{2}\sqrt{\frac{L}{C}}$$
$$\beta \simeq \omega\sqrt{LC} = \frac{\omega}{c} \qquad (1.7.)$$
$$\gamma = \alpha + j\beta = \alpha + j\frac{\omega}{c}$$
$$Z_0 \simeq \sqrt{\frac{L}{C}}$$

c : vitesse de transmission sur la ligne



$Za = Z_0 th(\gamma l/2)$ $\qquad$ $Ya = Y_0 th(\gamma l/2)$

$Zb = Z_0 coth(\gamma l)$ $\qquad$ $Yb = Y_0 coth(\gamma l)$

Figure (1.8.) : Quadripôle équivalent d'une ligne de transmission uniforme de longueur l.

(les constantes R,L,C,G sont définies par les caractéristiques du canal vocal /3/).

Nous supposerons, de plus , dans la suite que :

- la radiation par la bouche est négligeable .
- l'impédance de la source est très grande par rapport à l'impédance caractéristique du circuit.

Nous pouvons donc représenter le système vocal par le circuit de la figure (1.

L'équation (1.6.) pour $E_2 = 0$, $I_1 = Ug$, $I_2 = Um$ donne la fonction de transfert du système:

$$\frac{Um}{Ug} = \frac{1}{ch(\gamma l)} = \frac{Zb}{Za + Zb} \qquad (1.8.)$$

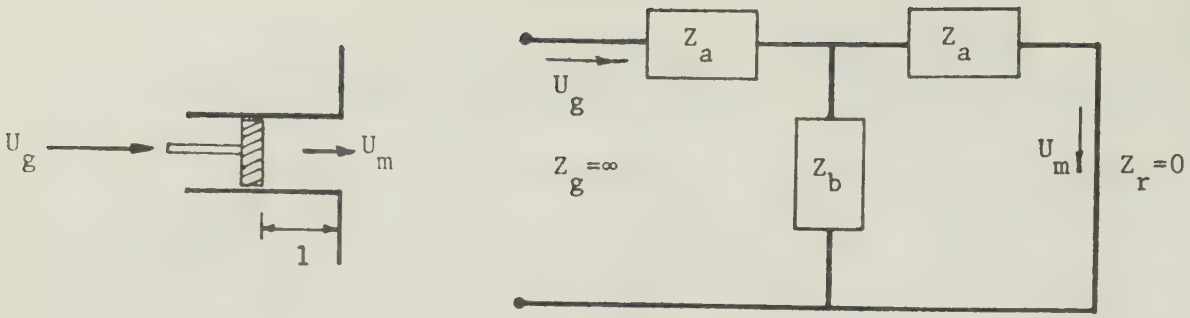


Figure 1.9. : Circuit équivalent du système vocal.

Les pôles de la fonction de transfert sont définis et calculés par l'équation :

$$ch\ (\gamma l) = 0$$

soit :

$$\gamma l = \pm j\ (2n + 1)\frac{\pi}{2} \qquad n = 0,\ 1,\ \ldots.$$

soit encore, après les simplifications (1.7.)

$$\alpha l + j\frac{\omega l}{c} = \pm j\ (2n + 1)\frac{\pi}{2}$$

A partir de cette équation , en posant $j\omega = p_n$, on prend les pôles

$$p_n = -\alpha c \pm j\frac{(2n + 1)\pi c}{2l} \qquad n = 0,1,\ldots. \tag{1.9.}$$

Remarquons qu'on peut également mettre l'équation (1.8.) sous la forme :

$$H(p) = \frac{Um(p)}{Ug(p)} = \prod_{n=1}^{\infty} \frac{p_n \cdot p_n^*}{(p-p_n).(p-p_n^*)} \tag{1.10.}$$

où p_n^* est le conjugué de p_n.

La transmission est alors caractérisée par une série infinie de pôles conjugués . Cherchons les fréquences pour lesquelles l'amplitude devient maximum.

Pour $\gamma = \alpha + j\beta = \alpha + j\frac{\omega}{c}$

$$|H(j\omega)| = \frac{2}{\{e^{2\alpha l} + e^{-2\alpha l} + 2\cos(2\omega l/c)\}^{\frac{1}{2}}}$$

alors :

$$\cos\{\frac{2\omega l}{c}\} = -1 \qquad \text{pour:}$$

$$\omega = \pm(2n+1)\pi c/2l \qquad n = 0,1,2,... \qquad (1.11.)$$

$|H(j\omega)|$ devient environ $\frac{1}{\alpha l}$ (α très petit)

On peut également calculer la bande passante à 3 db autour de chaque pic . On trouve alors :

$$\Delta\omega = 2\alpha c \ , \quad \Delta f = \alpha c/\pi \qquad (1.12.)$$

La figure (1.10.) donne l'allure en amplitude et phase de la fonction de transfert (1.10.) .

En conclusion ce modèle simplifié de production des sons met en évidence la structure du spectre du signal vocal ; les formants y apparaissent clairement . De plus l'équation(1.10.)montreque la transmission entre la glotte et la bouche peut être assimilée à une série de résonateurs connectés en cascade et dont chacun correspond à un formant du spectre. Il peut être entièrement défini par la fréquence du formant et la bande passante de celui-ci . Sur la base de l'équation(1.10.)il est possible de construire un modèle mathématique du système vocal , et c'est ce modèle que l'on examinera dans le chapitre suivant .

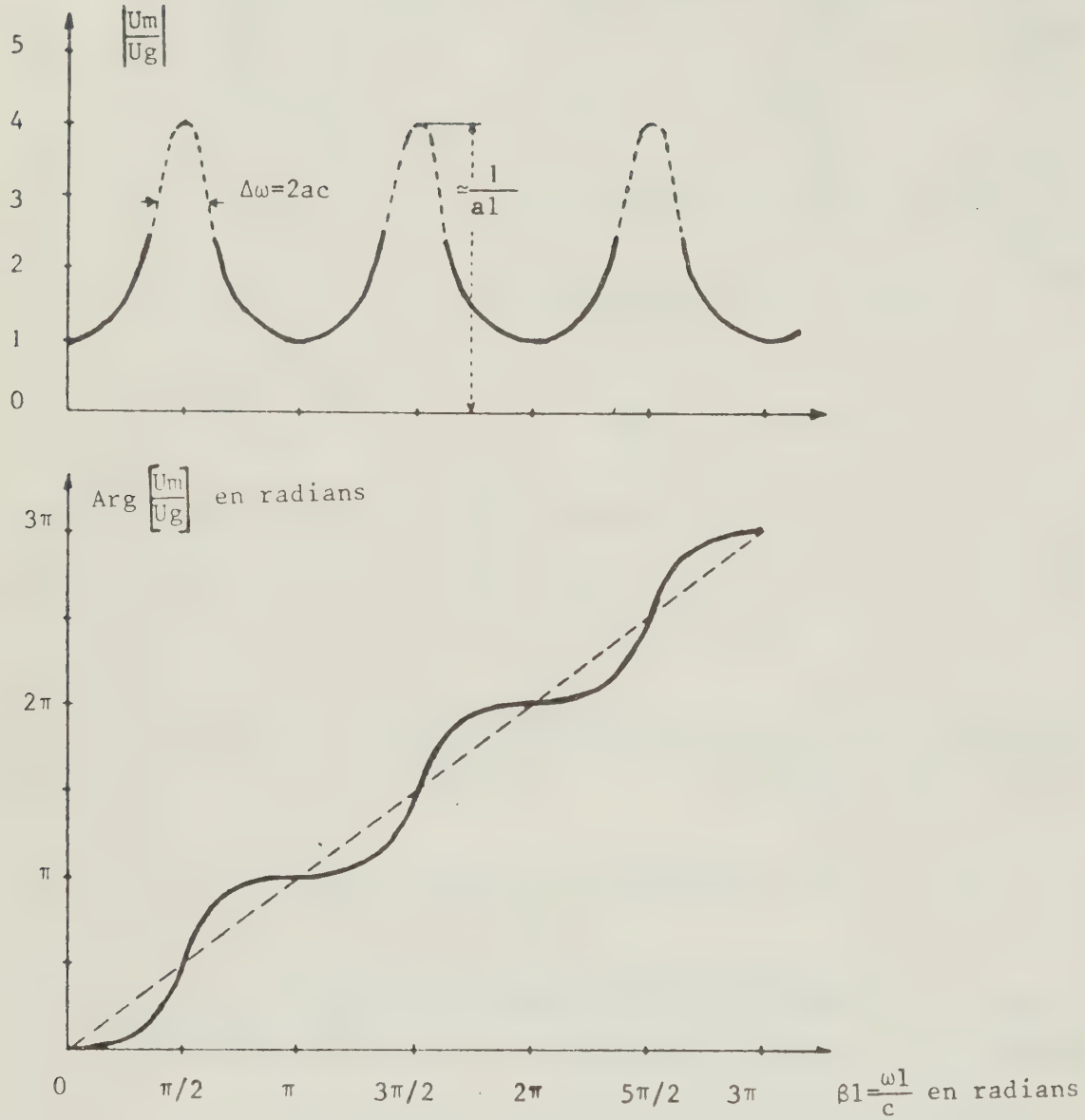


Figure 1.10 : Réponse en amplitude et phase du modèle simplifié du canal vocal.

1.6 Caractérisation mathématique des voyelles.

Les voyelles sont des sons sonores, où les cordes vocales constituent la seule source d'excitation du système vocal. Leur spectre représente une série de maxima , et comme l'a montré l'étude précédente, le système de transmission est caractérisé par une série de formants. On est amené , alors, à représenter l'appareil vocal, dans le cas de voyelles,

comme une cascade de systèmes du second ordre , ayant chacun deux pôles conjugués.

L'analyse précédente a montré que la fréquence du formant est la partie imaginaire du pôle tandis que la bande passante (à 3 db) est la moitié de la partie réelle.

Si Bn représente la bande passante du formant en Hz et Fn la fréquence du formant en Hz, on peut écrire l'équation(1.10.) comme suit:

$$P_n = \frac{Bn}{2} \pm jF_n \quad (HZ) \qquad (1.13$$

et on peut considérer l'équation (1.10) comme équation du système de transmission . Remarquons que les paramètres Bn et Fn peuvent être définis de façon expérimentale. L'équation (1.10) est encore simplifiée car ce sont les trois ou quatre premiers formants dont l'influence est primordiale dans le spectre, tandis que pour les autres , on introduit une fonction correctrice . Si l'on présente par H(f) le système de transmission de la glotte à la bouche , pour avoir le schéma du système global, il suffit d'ajouter la source d'excitation E(f) ainsi que la diffusion par la bouche R(f) . On obtient alors le système présenté par la figure (1.11.).

Figure 1.11. : Schéma-bloc du système de production de voyelles

On remarque que (voir /1/).

E(f) est presque proportionnelle à $1/f^2$, après une fréquence de 100 Hz, c'est à dire :

$$E(f) = \frac{k_v}{1+(f/100)^2} \qquad (k_v : \text{constante})$$

R(f) : est proportionnelle à la fréquence

$$R(f) = k_R.f \qquad (k_R : \text{constante})$$

Le système de transmission est tel que :

$$|H(f)| = \prod_{n=1}^{\infty} |H_n(f)| \qquad (1.14.)$$

et $$|H_n(f)| = \frac{p_n.p_n^*}{(p-p_n).(p-p_n^*)} \quad , \quad \text{pour} \quad p = jf \qquad (1.15.)$$

En considérant seulement les quatre premiers formants et en appelant $k_c(f)$ la fonction correctrice correspondante, c'est à dire celle qui tient compte de l'influence des formants supérieurs, on peut écrire :

$$|H(f)| = k_c(f).|H_1(f)|.|H_2(f)|.|H_3(f)|.|H_4(f)| \qquad (1.16.)$$

D'après l'équation (1.13.) l'équation (1.15.) devient:

$$|H_n(f)| = \frac{F_n^2 + (B_n/2)^2}{\{((f-F_n)^2 + (B_n/2)^2).((f+F_n)^2 + (B_n/2)^2)\}^{\frac{1}{2}}} \qquad (1.17.)$$

et puisque le rapport F_n / B_n est très grand dans le cas considéré , on peut écrire $Q = F_n / B_n$, Q étant le coefficient de surtension d'un simple résonateur représenté par les pôles p_n et p_n^*.

En posant $x = f/F$ l'équation (1.17.) devient :

$$|H_n(f)| \simeq \frac{1}{\{(1-x^2)^2 + x^2/Q^2\}^{\frac{1}{2}}} \qquad (1.18.)$$

Les figures 1.12. et 1.13.(reproduites de /1/) donnent une idée quantitative de l'influence de divers facteurs du système global.

En conclusion cette analyse fournit un modèle mathématique simple pour la représentation des voyelles . En ce qui concerne les autres sonss puisque d'une part la source est plus riche (bruit et cordes vocales éventuellement) et d'autre part le système de transmission possède de plus des zéros, le modèle mathématique devient plus compliqué . (pour une étude plus approfondie voir /1/ et /3/).

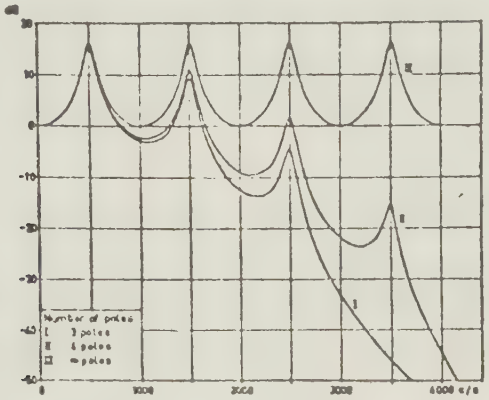


Figure 1.12. : Réponse en fréquences du système de transmission de la glotte à la bouche.

III . Modèle idéal, équation (10)
II . Approximation par 4 pôles
I . Approximation par 3 pôles

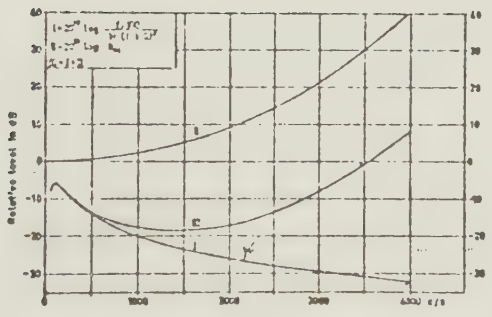


Figure 1.13. : facteurs de correction appliqués à la représentation des voyelles par un système à 4 pôles.

I. Source et radiation par la bouche
II. Contribution des pôles supérieurs .
III. Somme de I et II.

2. SIGNAL VOCAL - DIFFERENTS SUPPORTS D'INFORMATION

Nous allons examiner ici, d'une manière descriptive le signal vocal. Nous savons déjà que le message parlé a comme support des variations rapides de pression de l'air, qui se traduisent par un signal compliqué. La complexité du signal est dûe à des facteurs très différents dont nous énonçons certains :

- l'existence des entrées de nature différente (voir CHI) ainsi que la possibilité de co-existence de plusieurs entrées
- la quasi-périodicité du signal
- la non-stationnarité du signal.

Cette "complexité" du signal est à l'origine de certaines difficultés à l'analyse. Ceci se traduit par des exigences supplémentaires :

Non seulement il faut tenir compte de la non-stationnarité du signal qui se traduit à court terme par des "modèles différents du système", mais en plus il faut tenir compte à chaque fois des "modèles différents de l'entrée" (cette entrée étant en général encore une inconnue). En outre, la séparation des influences de l'entrée et du système proprement dit reste une opération assez délicate.

Nous avons déjà remarqué que le signal vocal était un signal très riche en informations diverses. Cherchons maintenant à savoir quelle est l'information utile contenue dans ce signal.

On peut trouver deux principaux types d'informations utiles :
1) Information sur "ce qui a été dit". Le terme "information sémantique" est souvent utilisé pour désigner ce type d'information. Cette information est liée aux"phonèmes" et à la succession des phonèmes pour former les mots et les phrases. Nous considérons que la notion de phonème est déjà rendu intuitive. En parlant dans le langage de la reconnaissance des formes, on pourrait définir les phonèmes comme tous les concepts "différents"relatifs au message parlé. Souvent le phonème est défini "comme l'unité linguistique de base" analogue au

caractère, unité de base du message écrit. Il correspond aux différents "sons" utilisés dans un language (voir CH1).
Nous préférons utiliser dans la suite le terme "Information phonémique" à la place de "sémantique".

2) Information sur la personne qui a parlé (locuteur). Cette information comprend la façon habituelle dont le locuteur s'exprime ainsi qu'une information "occasionnelle" liée à son état physique et à son état émotionnel.

On tachera, par la suite, de regrouper l'information contenue dans le message parlé à différents supports fondamentaux.
Ces supports sont les suivants :

- Variations d'amplitude du signal vocal (enveloppe énergétique)
- Variations du fondamental
- Régions stables d'un phonème
- Régions de transitions entre phonèmes
- Durée des états stables ou transitoires (information temporelle).

Ce regroupement est essentiel aussi bien pour la compréhension du mécanisme de la voix que pour l'évaluation de l'orientation de la recherche sur la parole dans le monde.

L'expérience a montré (CH 8) que l'intervention et la modification au niveau des variations de certains de ces supports, n'a pas affecté la qualité de la parole. Il serait donc intéressant de pouvoir juger de cet aspect. Nous allons par la suite examiner plus en détail chacun de ces facteurs informationnels.

1) Enveloppe énergétique : Il est facile d'accéder à ce support d'information. Il suffit de mesurer la puissance du signal dans des intervalles déterminés. L'enveloppe énergétique est très informative. Ses variations contiennent une grande partie de l'information concernant le locuteur : des expériences qui ont été entreprises ont démontré que la vérification d'un locuteur était possible en utilisant uniquement

l'enveloppe énergétique d'une phrase code avec un pourcentage de vérification correct de 95 %.
Pourtant la participation "phonémique" de l'enveloppe énergétique est réduite voir minime : des expériences de synthèse avec une enveloppe constante ont donné des phrases intelligibles.

2) Fréquence fondamentale : Il s'agit de la fréquence naturelle de vibration des cordes vocales comme ceci a été expliqué au Chapitre 1 : "l'analyse de la fréquence fondamentale", ou "analyse du pitch" a deux aspects différents sur le signal vocal :

a) l'aspect : mesure de la fréquence fondamentale
b) l'aspect : détection d'une nouvelle période.

On retrouve ces deux aspects sur les méthodes d'analyse. En général, l'aspect mesure est lié aux méthodes asynchrones d'analyse. Une mesure du pitch est effectuée sur un certain nombre de périodes. Ceci conduit la plupart du temps, à des valeurs moyennes du pitch sur un nombre de périodes. L'aspect détection est lié à une nécessité des méthodes synchrones, parce qu'on doit connaître exactement les instants de début de chaque période pour appliquer les algorithmes d'analyse, étant donné que la période peut être considérée comme un "processus unitaire stationnaire" qui se répète à court terme pour constituer le phonème sonore (voir fig 2.1).

Comme méthodes de mesure, on peut noter celle du cepstre (NØLL) /4/ et celle d'ITAKURA-SAITØ /5/ basée sur l'utilisation d'un modèle du canal vocal (voir chapitre 5).

Comme méthodes de détection, on note celle de GØLD /6/ basée sur la reconnaissance des formes, celle de MAISSIS /7/ basée sur les variations énergétiques du signal vocal et celle du GUEGUEN - CARAYANNIS basée sur l'utilisation du filtre de KALMAN pour l'analyse de la parole (voir chapitre 7).

Il faut ajouter que beaucoup de recherches ont été consacrées depuis longtemps à la mesure du pitch, mais le problème ne semble pas être résolu complètement. En effet, dans certaines conditions cette mesure ne semble pas être facile sur le signal vocal.

Le caractère informationnel des variations du pitch est le même que celui de l'enveloppe énergétique. Il ne semble pas que l'information phonémique soit sensible à ces variations du fondamental. Des expériences de synthèse faites avec une fréquence fondamentale constante ont donné des formes intelligibles /8/.

Par contre, une grande partie de l'information relative au locuteur est liée aux variations du fondamental au cours du temps (/31/, /32/). Nous donnons ici (voir tableau 2), les résultats des expériences de vérification de locuteurs publiés par LUMMIS /9/. Une phrase code est utilisée. A gauche du tableau, on donne les informations utilisées chaque fois, à droite le pourcentage d'erreur de vérification.

3-4) Régions stables et régions transitoires d'un phonème :

La plus grande partie de la recherche sur l'analyse s'oriente de façon à pouvoir représenter ces régions d'une manière pertinente. Dans le présent rapport de travail, on peut trouver un grand nombre de méthodes numériques dans le but d'une telle représentation. Le terme représentation a ici une double signification :

- 1) Pertinence
- 2) Diminution de la redondance

C'est dans ce sens que toutes les méthodes d'analyse ont été développées, vu que toute application (reconnaissance, transmission des données, identification des locuteurs) se heurte au grand volume des données.

Les régions stables et transitoires des phonèmes contiennent l'information phonémique. Concernant les régions transitoires, l'évolution est considérée, en général, comme assez lente, pour qu'on puisse représenter ces régions par une succession de segments considerés comme approximativement stationnaires.

Informations Utilisées	Erreur de verification
Enveloppe énergétique	4.1 %
Pitch	3.2 %
Pitch + Enveloppe énergetique	1.4 %
Pitch + enveloppe énergetique + ajustement temporaire	1.2 %
Pitch + enveloppe énergetique + ajustement temporaire + variations des formants.	1.0 %

TABLEAU 2 : Erreur de verification de locuteur en fonction des informations utilisées. (D' après LUMMIS /9/)

5) Durée des états stables et transitoires

A cette durée est liée la très importante notion de segmentation, c'est-à-dire la possibilité de décision relative aux frontières des phonèmes. Le problème se présente de façon analogue à celle de la détection d'une nouvelle période du pitch. Le nouvel évènement est ici le nouveau phonème.

Nous présentons dans ce rapport au chapitre 9 une méthode de segmentation automatique. En outre, un modèle est fourni pour la transition (mode de passage, ou trajectoire). La "détection" se fait avec une décision du type "non appartenance" à une classe déjà identifiée (modèle relatif au processus générateur de bruit sur la région stable).

La durée des états stables et transitoires est une information qui est liée au locuteur (rapidité de la locution) mais souvent occasionnellement. Donc ce n'est pas une information qui pourrait caractériser le locuteur d'une manière pertinente. C'est la raison pour laquelle un ajustement temporel des paramètres relatifs est souvent exigé.

Nous donnons par la suite (figures 2.1 - 2.12), un certain nombre de signaux de la parole pour familiariser le lecteur avec l'aspect temporel de ce signal.

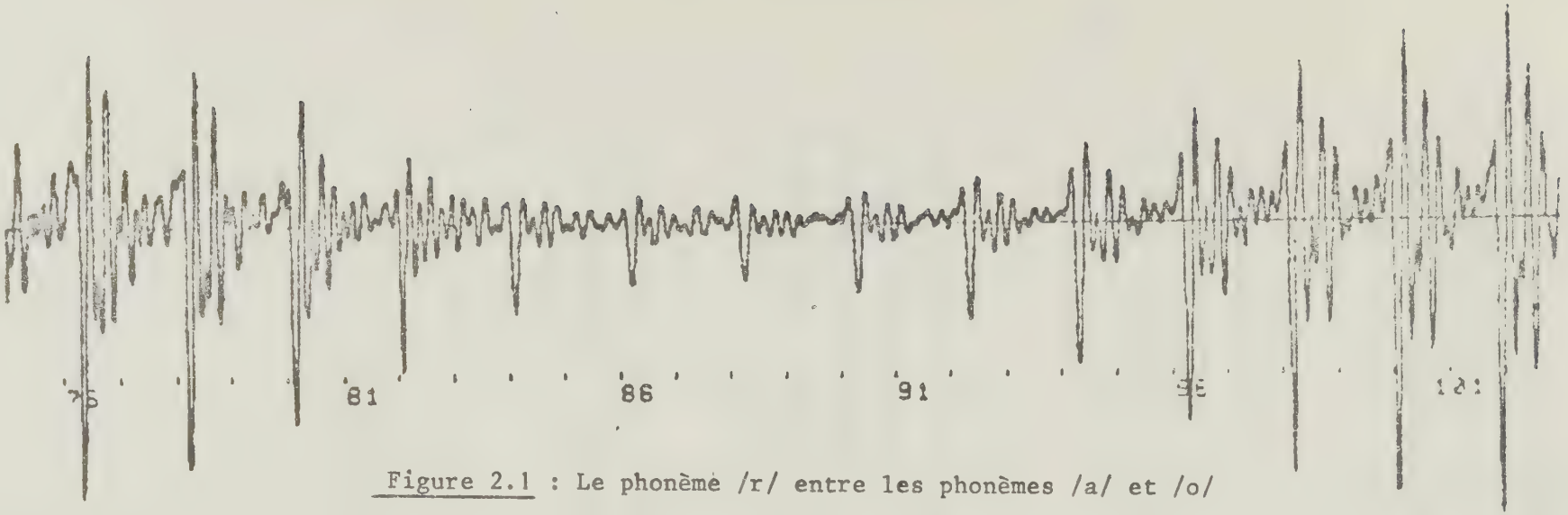


Figure 2.1 : Le phonème /r/ entre les phonèmes /a/ et /o/

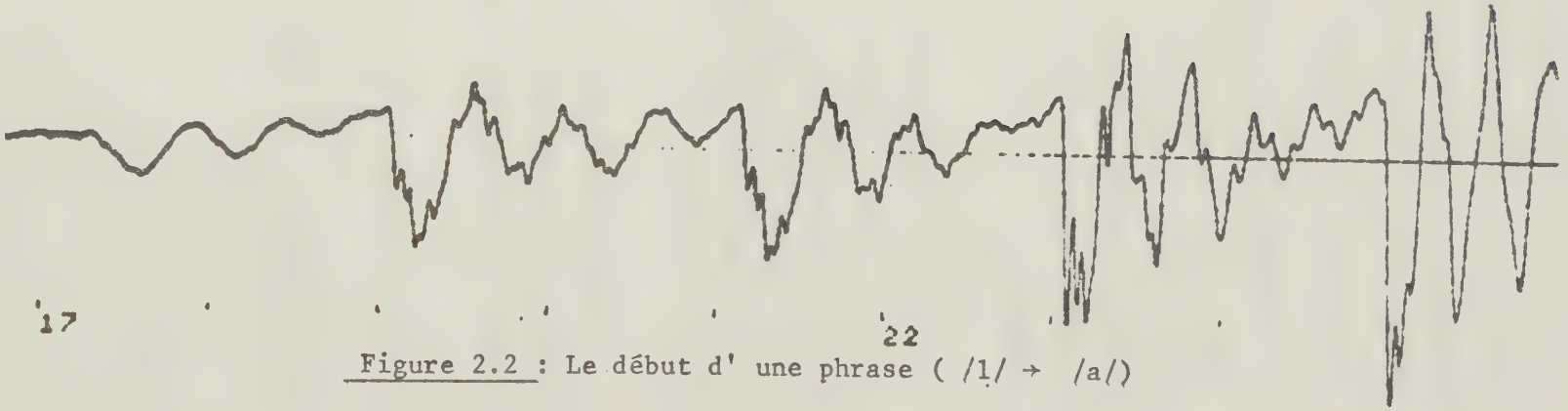


Figure 2.2 : Le début d' une phrase (/l/ → /a/)

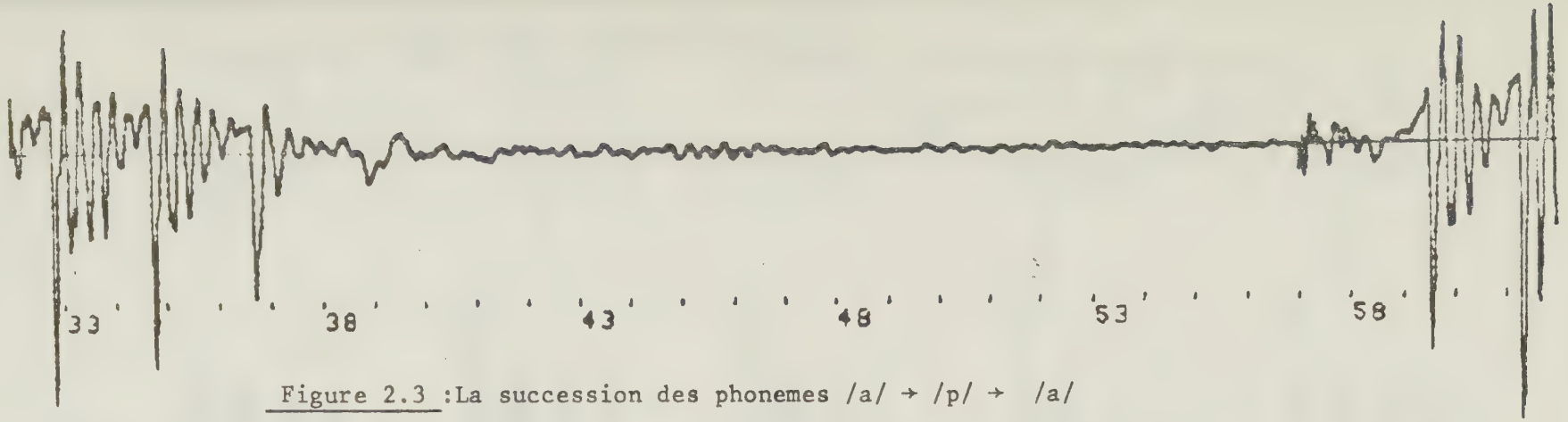


Figure 2.3 :La succession des phonemes /a/ → /p/ → /a/

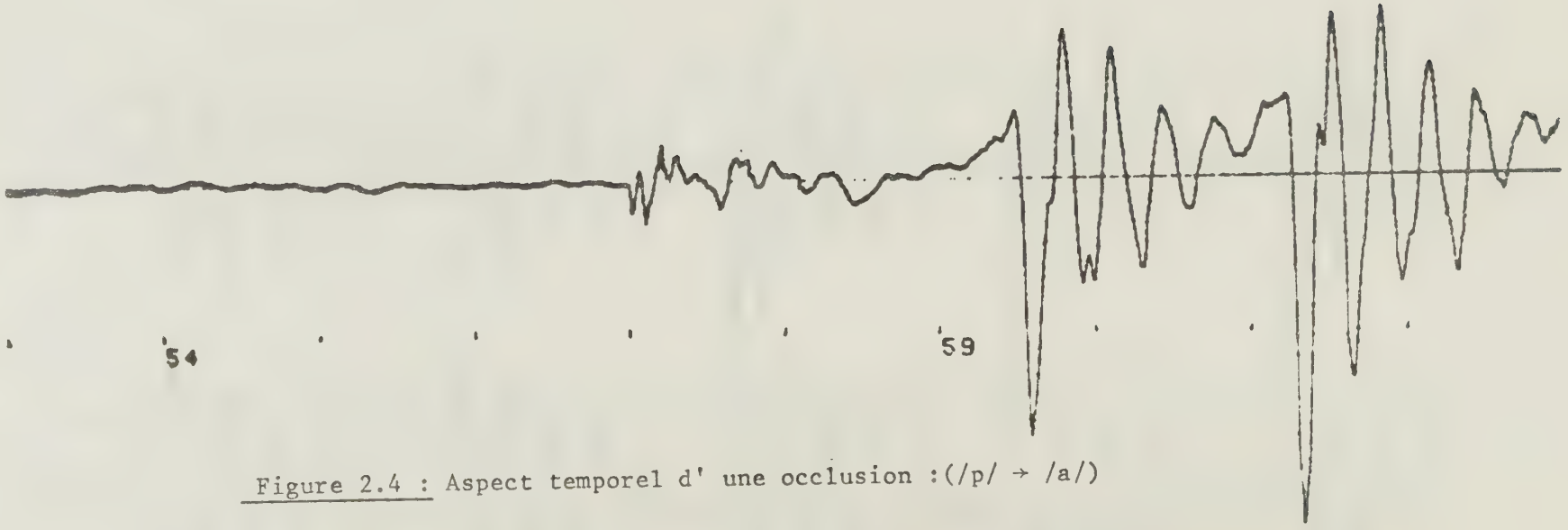


Figure 2.4 : Aspect temporel d' une occlusion :(/p/ → /a/)

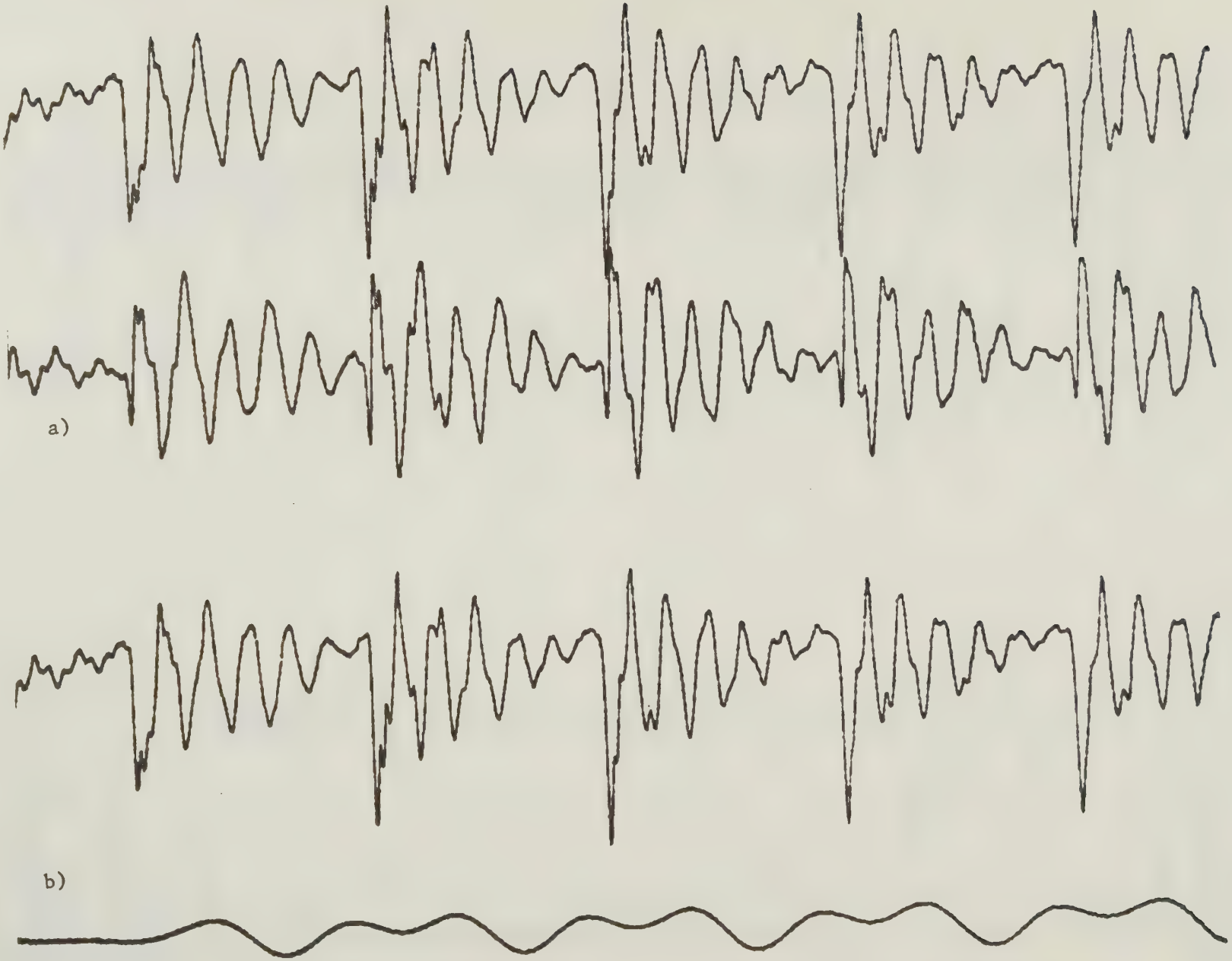


Figure 2.5 : Signal vocal filtré: a) passe-haut , b) passe-bas. Frequence de coupure 200 Hz dans les deux cas .

Figure 2.6 : Signal vocal filtré : a) passe-haut , b) passe bas . Frequence de coupure 500 Hz dans les deux cas . Nous constatons un dephasage du signal (b) , dû au filtrage. Ce dephasage empêche la détection du pitch sur ce signal

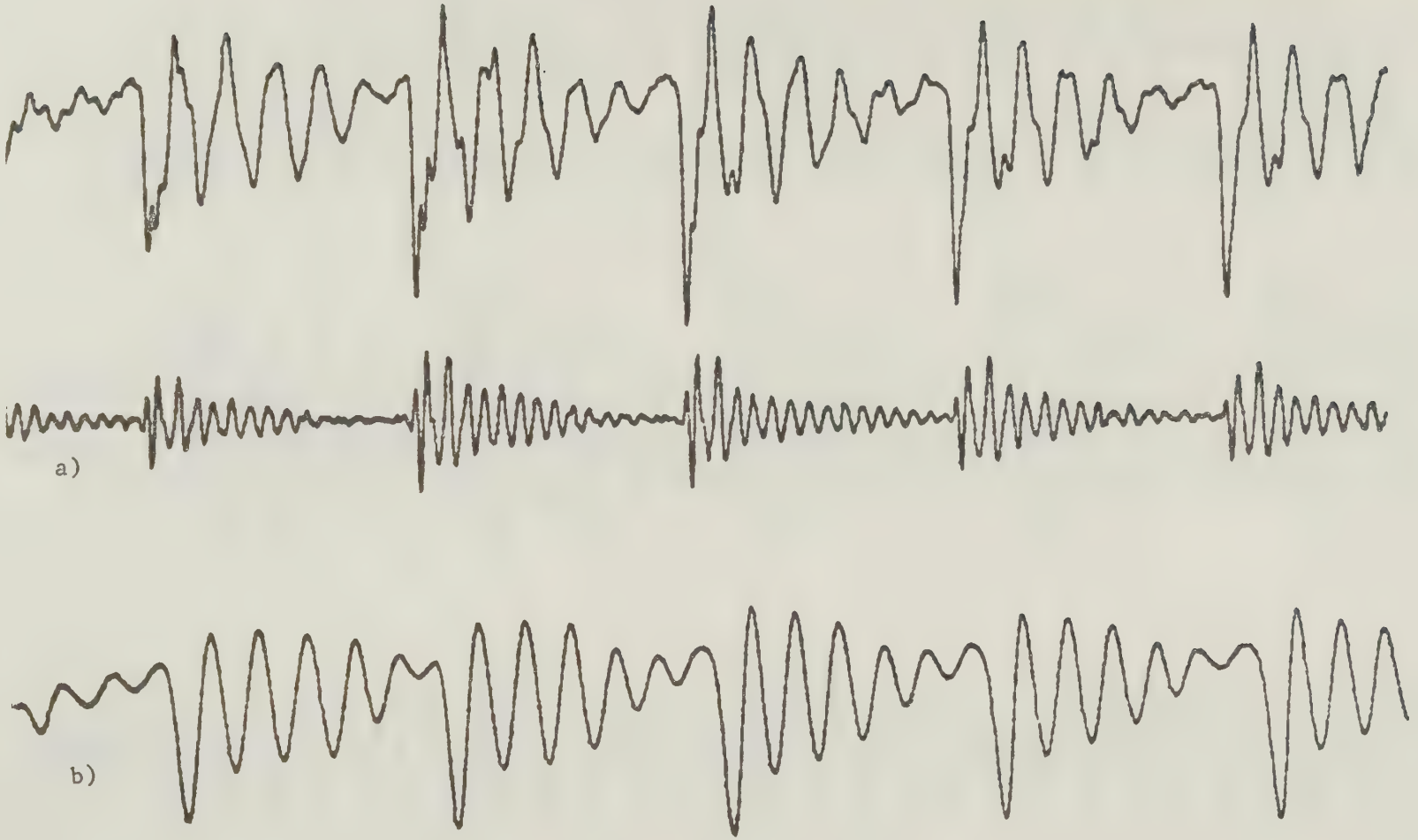


Figure 2.7 : Signal vocal filtré : a) passe-haut , b) passe-bas. Frequence de coupure 1000 Hz

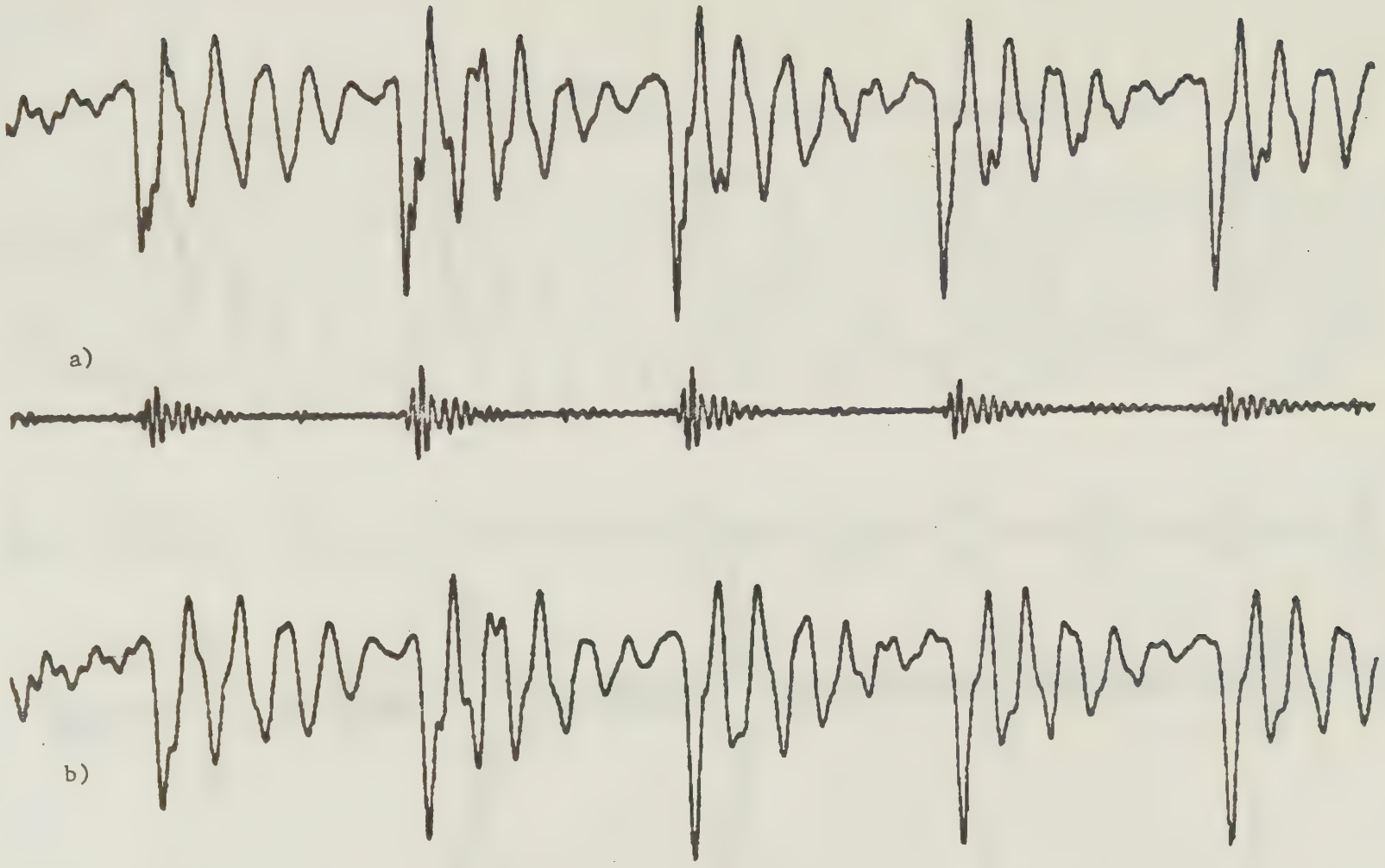


Figure 2.8 : Signal vocal filtré : a) passe-haut , b) passe-bas. Frequence de coupure 2000 Hz

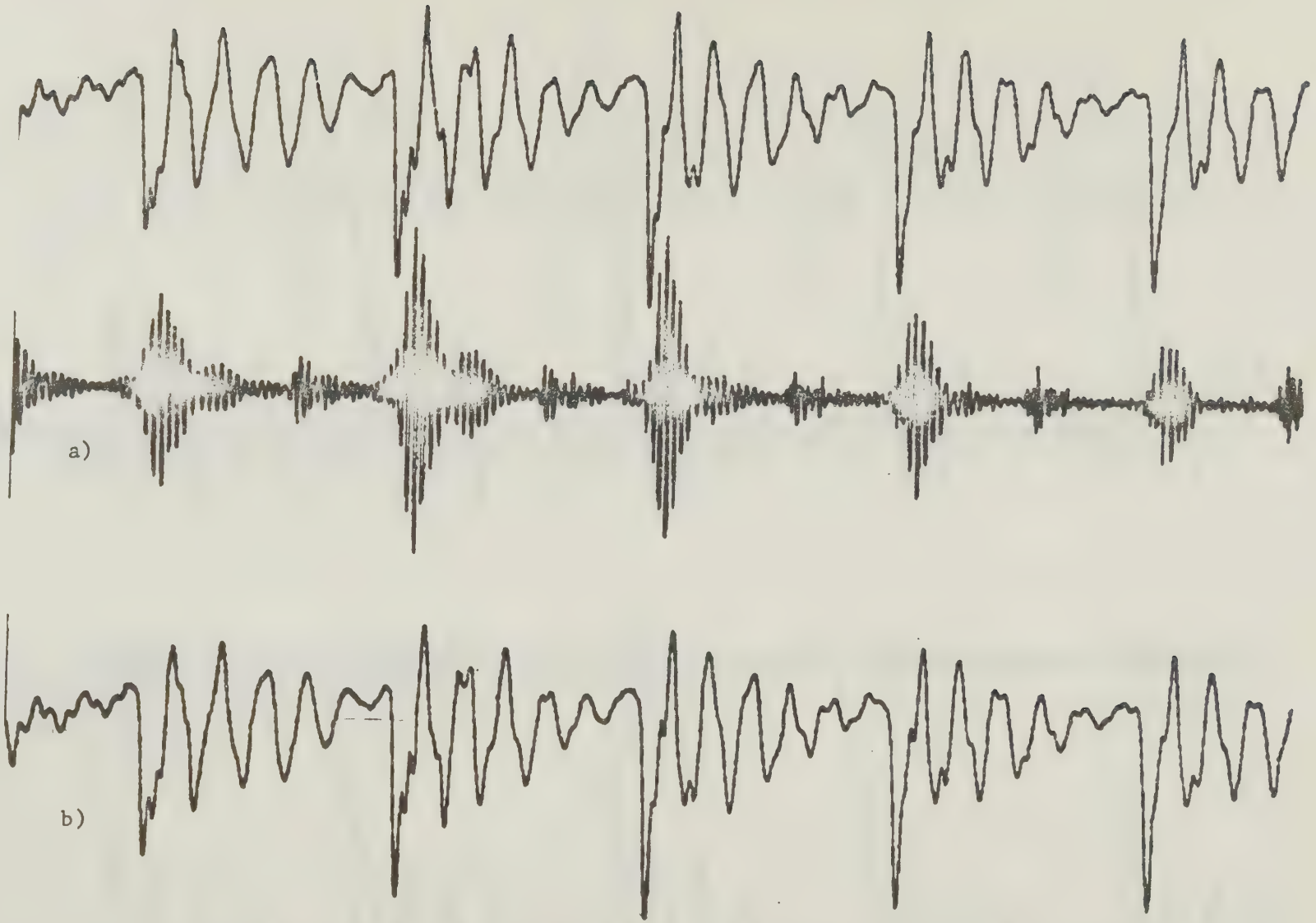


Figure 2.9 : Signal vocal filtré : a) passe-haut, b) passe-bas . Frequence de coupure 3000 Hz

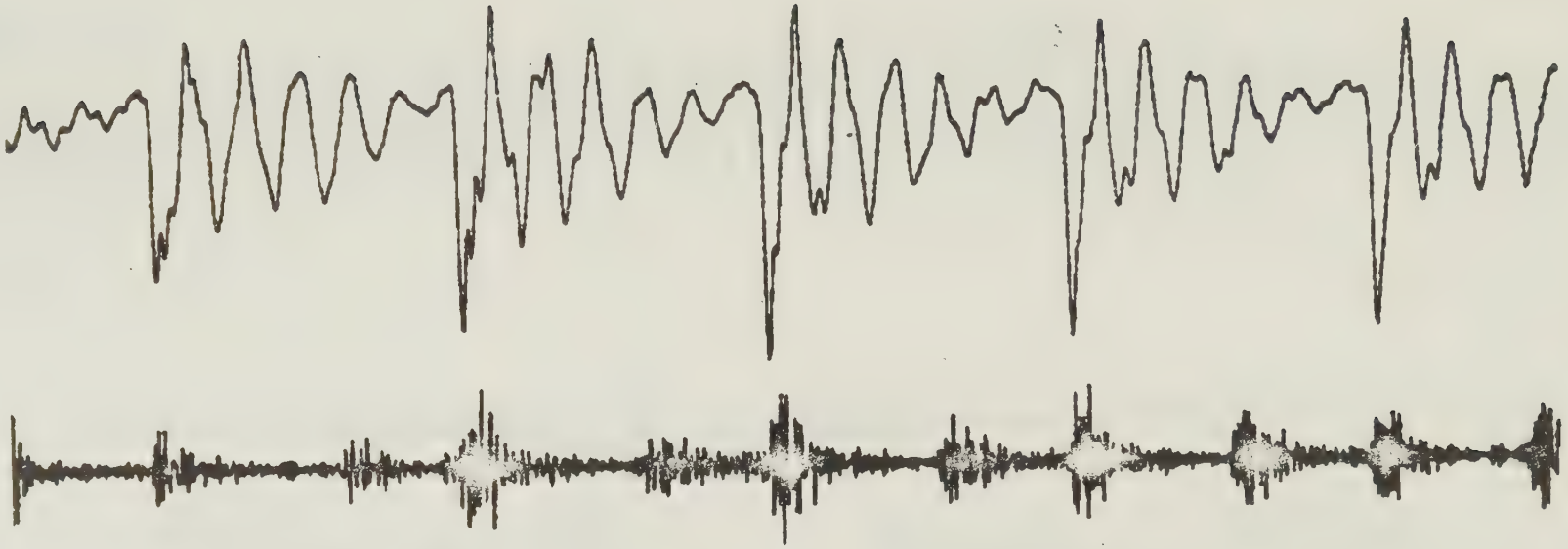


Figure 2.10 : Signal vocal filtré avec un filtre passe-haut. Frequence de coupure 4000 Hz.

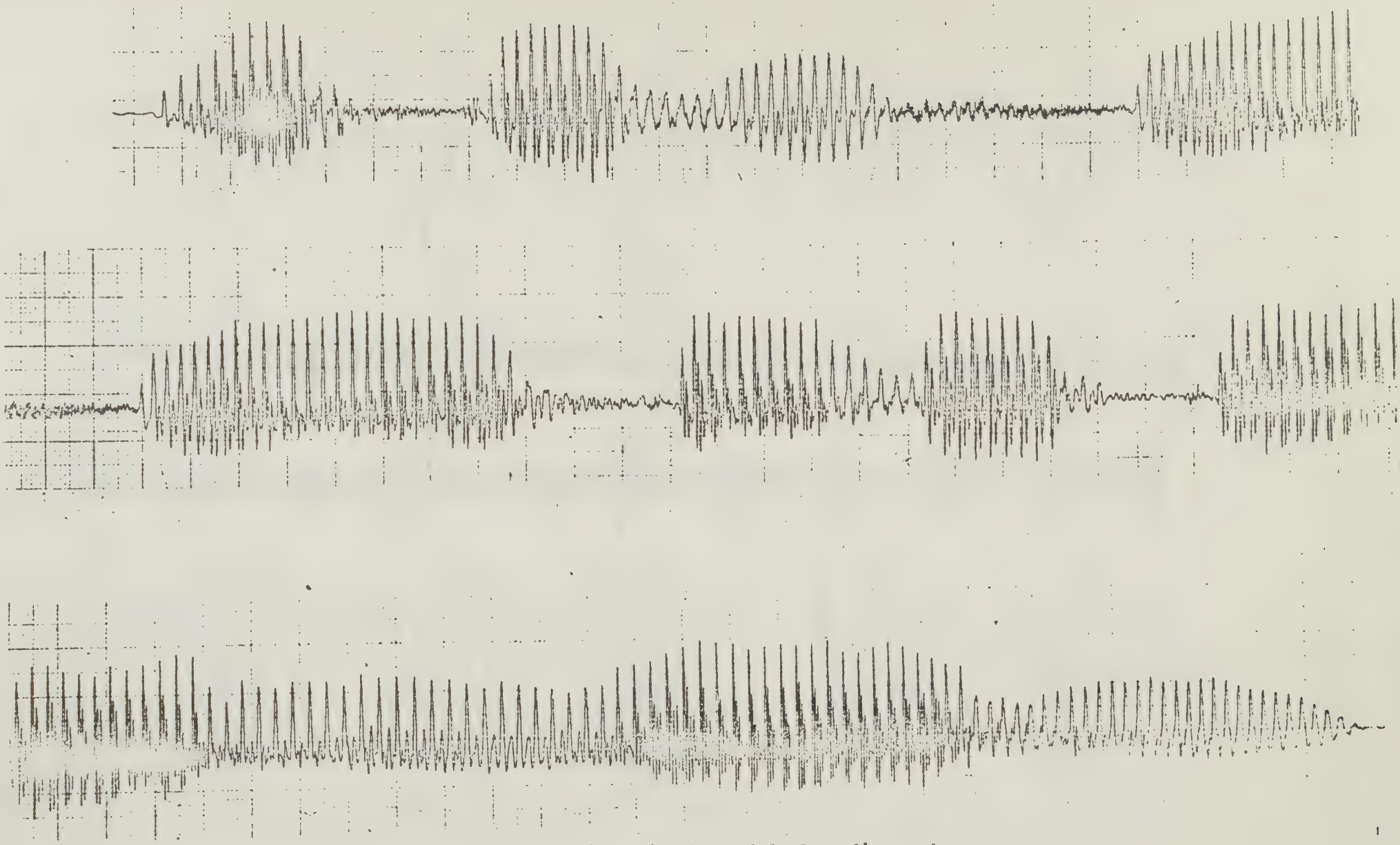


Figure 2.11 : Evolution d' un signal vocal le long d' une phrase.

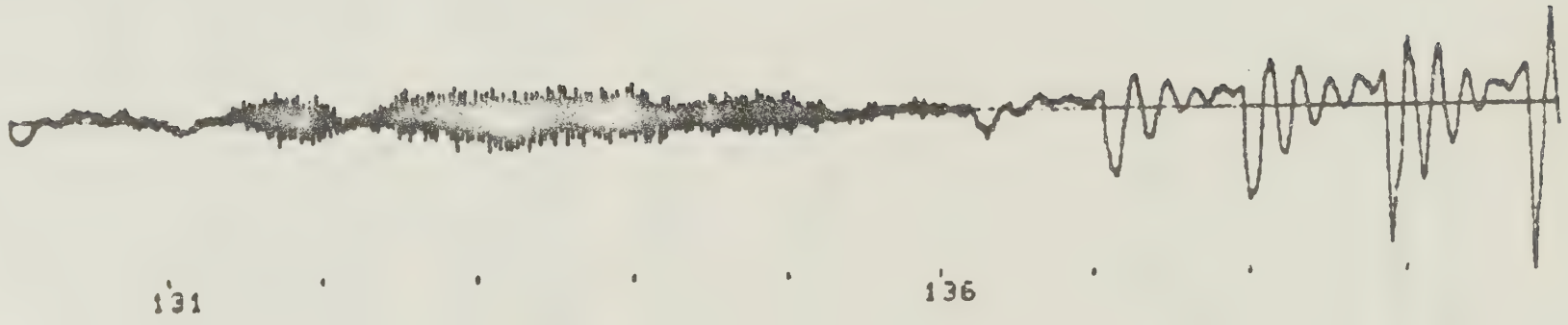


Figure 2.12 : Succession d'un son sourd à un son sonore (/g/ --→ /an/)

3. ANALYSE ET PARAMETRISATION DE LA PAROLE - METHODES CLASSIQUES D'ANALYSE

3.1. Définitions

Le but d'un système d'analyse est d'extraire un nombre suffisamment restreint de paramètres caractérisant le signal de manière pertinente. Ce souci répond à la nécessité d'une réduction de la redondance de la parole destinée à garder seulement l'information utile. La qualité de ce traitement est en effet essentielle dans tout système de synthèse vocale, de reconnaissance automatique de la parole ou d'identification des locuteurs, où la mise en oeuvre des algorithmes se heurte au volume prohibitif des données brutes.

En général, le mot "analyse" sera associé à des paramètres qui peuvent donner lieu à une synthèse ou à une reconstitution du signal analysé. Par contre, le mot "paramétrisation" sera réservé pour une analyse après laquelle la reconstitution est impossible.
Donc, avec le vocable "analyse", deux choses seront sous-entendues :

1. Diminution de la redondance
2. Synthèse (reconstitution) possible.

Le vocable "représentation" de la parole est souvent utilisé pour caractériser le résultat de cette analyse. L'expression "codage de la parole "speech coding" est aussi utilisé dans le même sens.

Un vocoder est un appareil qui effectue un tel codage. Il est, en général, constitué de deux parties principales : une partie qui effectue l'analyse du signal et qui est attachée à l'émetteur et une autre qui effectue la synthèse et qui est attachée au récepteur (voir figure 3.1 et 3.3).

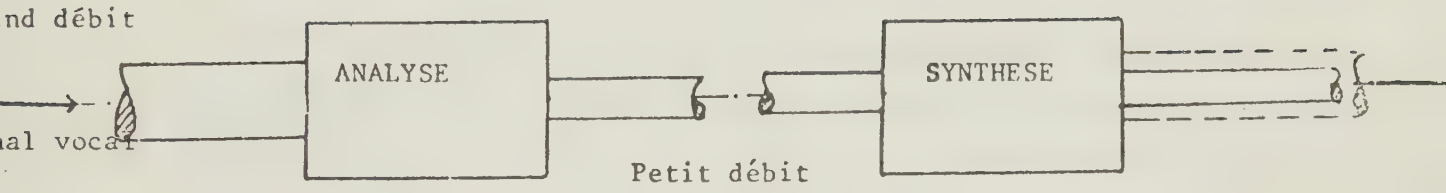


Figure 3.1. Principe d'un vocoder

Réduction de la redondance de la parole en vue d'une transmission économique .

En vue d'une application particulière , certaines qualités sont de dées à un algorithme . En vue d'une autre application des qualités supplémentaires ou peut être des qualités tout à fait différentes peuve être exigées . Donc pour parler des qualités il faut se placer dans le cadre d'une certaine application .

Par exemple des applications comme la reconnaissance de la parole mais surtout la transmission des données , doivent être attachées à un analyseur fonctionnant en temps réel . En outre dans le cadre de la transmission économique des données , il faut en plus (dans le but d'un taux de transmission réduit et d'une économie en espace mémoire)

a) que le nombre des paramètres soit relativement restreint
b) que les paramètres ne soient pas très sensibles aux erreurs de quantification .

Parmi les méthodes d'analyse existantes, certaines peuvent répondre à un grand nombre d'exigences mais aucune ne peut répondre à toutes .

Des méthodes rapides , où l'implantation en temps réel est possible , apportent peu d'information ; des méthodes apportant beaucoup d'information étant en général plus lentes : l'implantation en temps réel se heurte donc au volume important des calculs .

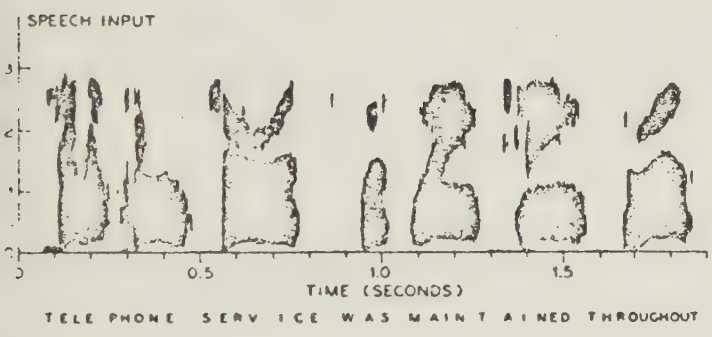
Deux types de méthodes existent pour analyser la parole . Des méthodes analogiques et des méthodes numériques .
Les méthodes analogiques peuvent répondre au souci du temps réel , mais la recherche de l'information n'est pas systématique . Avec les méthodes numériques la recherche est systématisée , mais le temps de calcul demandé est souvent important. De cette façon l'objectif du temps réel peut difficilement être atteint . (Sauf avec des appareils très spécialisés).

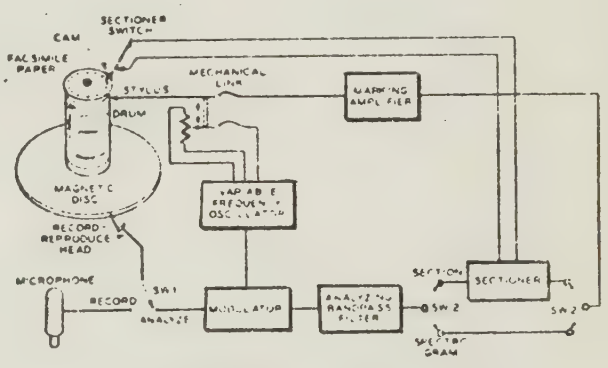
3.2. Méthodes Analogiques.

Les méthodes analogiques sont les méthodes les plus anciennes de l'analyse de la parole . Il s'agit déjà de méthodes très classiques dont on connait bien les limites , et qui sont bien expliquées dans la littérature (/1/ , /3/)
Ce travail est plutôt orienté vers les méthodes numériques . On se contentera seulement ici de présenter brièvement les principes des méthodes analogiques pour pouvoir ensuite placer les méthodes numériques dans un contexte général.

Ce sont surtout des analyseurs basés sur la définition des spectres et du spectre à court terme (voir chapitre 3.3.2.1.) qui ont été utilisés. On peuttrouver deux sortes d'analyseurs , des analyseurs hétérodynes et des analyseurs ayant une batterie de filtres passe-bande.

L'analyseur hétérodyne mélange le signal vocal avec une sinusoïde produite par un oscillateur . La différence des fréquences qui en résulte est appliquée sur un filtre passe-bande pour choisir la bande de fréquence qui contient la composante à déterminer . En réalité le signal d'entrée est balayé en fréquence avec un filtre de fréquence fixe. Cette méthode peut fournir des spectres avec une bonne résolution mais elle est très lente . Elle a trouvée des applications surtout dans le cadre sonographe . (Fig 3.2)





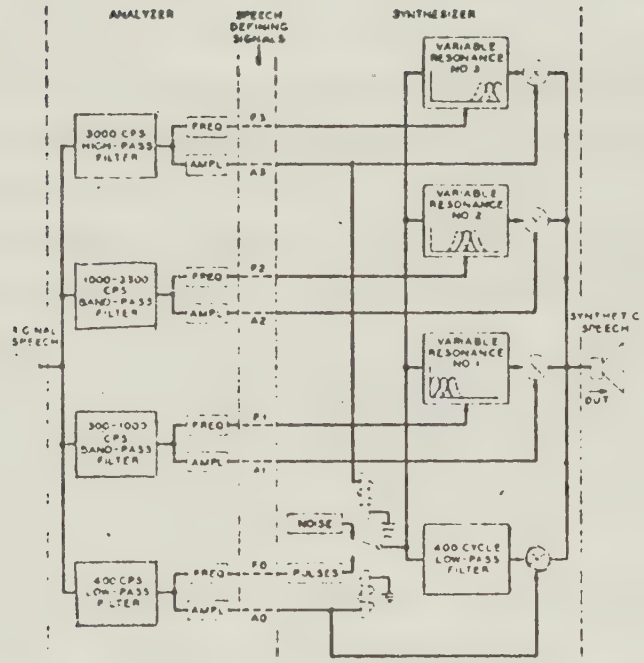
Functional diagram of the sound spectrograph

Figure 3.2. : Sonographe et sonogramme.

(figure tirée de /3/)

L'autre type d'analyseur est composé d'une série de filtres passe-bande . la sortie de chaque filtre correspond à une composante du spectre . Le spectre produit n'a pas une bonne résolution sauf si le nombre de filtres est très important , mais l'avantage de cette méthode est qu'elle peut être implantée en temps réel par définition (sortie instantanée) .

Sur cette technique sont basés les vocoders classiques(figure 3.3.)Dans ce genre d'appareil l'analyse se fait d'une manière analogique mais la transmission est digitale .



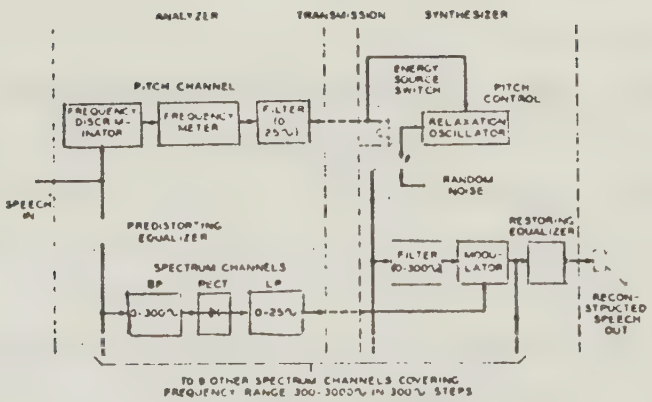


Figure 3.3 : Vocoder à formants et vocoder à canaux (figures tirées de /3/)

3.3. Méthodes Numériques .

3.3.1 Généralités .

Ces méthodes sont mises en oeuvre en utilisant un programme de traitement numérique et ne doivent en aucun cas être vues comme des simulations d'un système analogique . En outre plusieurs de ces techniques ne peuvent se réaliser avec un équipement analogique . Pourtant, vu la correspondance biunivoque programme-machine et l'évolution actuelle de la technologie digitale, ces techniques peuvent être réalisées avec des machines spécialisées complètes. Dans le cadre de ces méthodes un modèle discret de production du signal vocal est plutôt retenu. Par là, on entend que le signal vocal après échantillonnage est considéré comme un processus intrinsèquement discret.

Une méthode d'analyse est appelée synchrone si le traitement est effectué avec la même cadence que l'arrivée d'une nouvelle impulsion du fondamental. Sinon la méthode est appelée asynchrone. Dans ce cas, le segment de traitement est choisi arbitrairement

Deux grandes classes de méthodes numériques existent : les méthodes fréquentielles et les méthodes temporelles . Les méthodes fréquentielles travaillent après passage dans le domaine fréquentiel . La recherche systèmatique de l'information se fait alors dans ce domaine . Le passage dans le domaine fréquentiel se fait en utilisant la transformée de Fourier discrète qui est un cas particulier de la transformée en z . Les méthodes temporelles traitent directement le signal vocal sans faire appel au domaine fréquentiel .

3.3.2. Analyse en fréquence.

3.3.2.1. Définitions - Spectre à court terme - Spectre évolutif.

Le spectre de puissance d'une fonction de temps s(t) peut être défini par :

$$P(\omega) = \left| \int_{-\infty}^{+\infty} s(t) \exp(-j\omega t)\, dt \right|^2 \qquad (3.1)$$

Soit en utilisant la fonction d'autocorrelation v(t) :

$$P(\omega) = \int_{-\infty}^{+\infty} v(\tau) \exp(-j\omega\tau)\, d\tau \qquad (3.2)$$

où $$v(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{+T} s(t)\, s(t+\tau)\, dt \qquad (3.3)$$

Les relations précédentes s'écrivent pour un signal apériodique :

$$v(\tau) = \int_{-\infty}^{+\infty} s(t)\, s(t+\tau)\, dt \qquad (3.4)$$

et $$P(\omega) = \int_{-\infty}^{+\infty} v(\tau) \exp(-j\omega\tau)\, d\tau \qquad (3.5)$$

où $$P(\omega) = S(\omega).\overset{*}{S}(\omega) \qquad (3.6)$$

et $$S(\omega) = \int_{-\infty}^{+\infty} s(t) \exp(-j\omega t)\, dt \qquad (3.7)$$

et $$v(0) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} P(\omega)\, d\omega \quad = \text{l'énergie totale du signal} \qquad (3.8)$$

Ces relations ne sont pas utilisables dans la pratique pour deux raisons:

-elles demandent une intégration dans un intervalle infini.
-elles supposent une connaissance de l'avenir.

Dans les cas réels nous pouvons seulement éffectuer ces intégrations jusqu'à l'instant actuel (t). Un spectre de puissance à court terme est donc défini comme étant l'intégrale:

$$P(\omega,t) = \left| \int_{-\infty}^{t} w(t-\tau)\, s(\tau)\, e^{-j\omega\tau}\, d\tau \right|^2 \qquad (3.9.)$$

La fonction $w(t - \tau)$ peut être vue comme une fenêtre temporelle sous laquelle passe le signal d'entrée. En outre dans la pratique on est souvent amené à traiter des signaux qui évoluent avec le temps, dont le spectre évolue. Cette définition n'étant donc pas toujours satisfaisante, la fenêtre w est définie différente de zéro seulement dans un intervalle $|\tau| \leq \tau_m$. Nous avons donc dans la pratique des signaux temporels limités entre $-\tau_m$ et $+\tau_m$ c'est à dire des signaux limités dans le temps. Ces signaux ont par conséquent des spectres limités dans le temps mais non limités en fréquences. (/4/; /8/).

Il serait intéressant de se placer par la suite dans le cadre d'un signal échantillonné.
La transformée de Fourier peut être vue comme un cas particulier de la transformée en z. Ainsi, une correspondance est établie entre le cas analogique, (transformée de Laplace - transformée de Fourier), et le cas discret, (transformée en z - transformée de Fourier discrète).
La transformée en z d'une séquence s_n est définie par :

$$S(z) = \sum_{n=-\infty}^{-\infty} s_n\, z^{-n} \qquad (3.10)$$

où s_n et $S(z)$ peuvent être des nombres complexes en général.

Cette équation peut être écrite pour un nombre fini d'échantillons N.

$$S(z) = \sum_{n=0}^{N-1} s_n\, z^{-n} \qquad (3.11.)$$

où la somme converge $\forall z \neq 0$

On peut aboutir à cette relation à partir de la transformée de Laplace d'une série d'impulsions de Dirac. En effet:

$$\mathcal{L}\left\{\sum_{n=0}^{N-1} s_n \delta(t - nT)\right\} = \sum_{n=0}^{N-1} s_n e^{-snT} = \sum_{n=0}^{N-1} s_n z^{-n} \qquad (3.12)$$

où $z = e^{sT}$

La relation $z = e^{sT}$ établit une correspondance entre le plan de Laplace ($s = \sigma + j\omega$) et le plan des z ($z = a + jb$), (voir Fig. 3.4.)

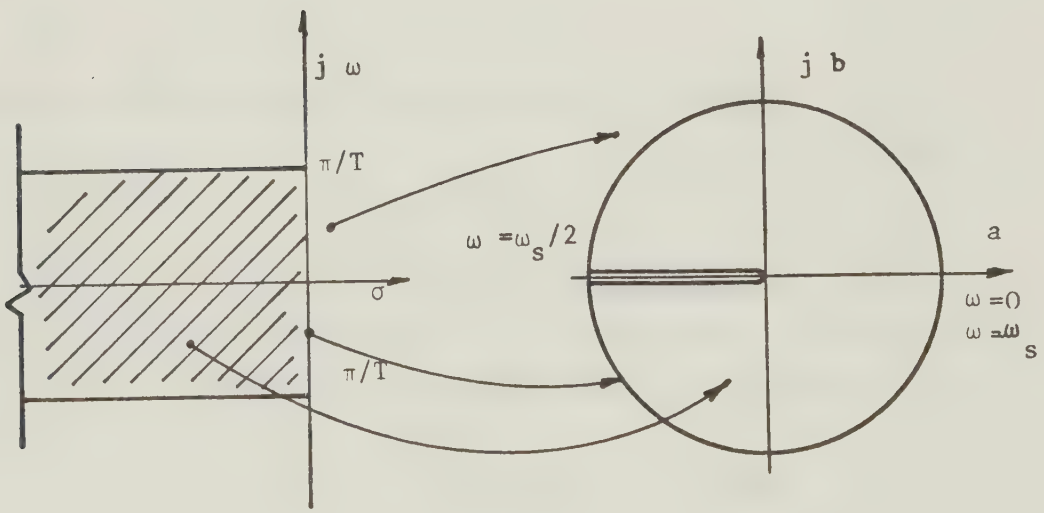


Figure 3.4. : Plan de Laplace et plan des z.

Pour une série d'impulsions la correspondance est telle que la région: ($-\infty<\sigma<+\infty$, $\pi/T \leq \omega \leq \pi/T$) corresponde au plan entier des z. L'axe jω correspond à la circonférence du cercle unitaire sur le plan des z. Nous savons que les valeurs de la transformée de Laplace sont souvent calculées sur l'axe jω ce qui donne lieu

à la transformée de Fourier . L'équivalent sur le plan des z serait le calcul de certaines valeurs sur la périphérie du cercle unitaire . Ceci donne la transformée de Fourier discrète . L'équation (3.4.) calculée sur la circonférence du cercle unitaire donnerait donc :

$$S_k = S(z_k) = \sum_{n=0}^{N-1} s_n \, z_k^{-n} \qquad (3.13)$$

$$\text{où} \quad z_k = \exp(j \, 2\pi k \, \delta f \, T)$$

où δf est un pas en fréquence et T la période d'échantillonage . L'équation (3.13.) est la définition de la transformée de Fourier discrète . Il s'agit d'une transformée de Fourier à court terme par définition . L'équation (3.13) suppose déjà l'existence d'une fenêtre temporelle rectangulaire sous laquelle le signal est vu. Une fenêtre différente peut être introduite d'une manière tout à fait naturelle dans cette équation . Il suffit de considérer au lieu de s_n des valeurs

$$s'_n = s_n \, w_n \qquad (3.14.)$$

Il faut noter que w_n pourrait être considéré comme la réponse impulsionnelle d'un système , excité par le signal s_n . Un pas égal à

$$\delta f = \frac{1}{N \, T} \qquad (3.15.)$$

peut être choisi pour le calcul du spectre . L'équation (3.13.) s'écrit dans ce cas :

$$S_k = \sum_{n=0}^{N-1} s_n \, z_k^{-n} \qquad \text{où} \quad z_k = \exp (j \, 2\pi k/N)$$

Soit

$$S_k = \sum_{n=0}^{N-1} s_n \exp(-j\ 2\pi\ nk/N) \qquad (3.16)$$

$$k = 1, \ldots, N$$

Ceci rend le calcul indépendant de la fréquence d'échantillonage , tout en lui donnant les moyens d'être effectué d'une manière plus rapide . En effet les symétries de l'équations (3.16.) peuvent donner lieu à une diminution du nombre d'opérations nécessaires pour le calcul du spectre .Si N est le nombre total des points considérés , au lieu d'un nombre de N^2 multiplications , $(N \log_2 N)$ multiplications sont nécessaires si N est une puissance de 2 . Sinon $(N \sum_i m_i)$ opérations sont nécessaires , où m_i sont les facteurs premiers de N . L'algorithme qui effectue ce calcul est connu sous le nom de la FFT (Fast Fourier Transform) /11/,/12/
La FFT a permis la construction de processeurs spéciaux (FFT processors) qui peuvent donner lieu à un calcul de spectre en temps réel même pour un signal relatif à la parole (bande passante utile 5 KHz)
Il est intéressant de noter que seulement la moitié des points calculés avec la FFT sont dans le domaine utile du spectre , le reste ne respectant pas le théorème de Nyquist .

Sur la figure (3.5.) on voit différents spectres à court terme calculés sous différentes fenêtres avec la FFT . L'amplitude du spectre est uniquement visualisée . On peut remarquer que les spectres calculés sous une fenêtre asses large contiennnent deux composantes , une composante qui " varie rapidement " et une composante " lente " . La première composante traduit l'influence de la source quasipériodique, la second traduit la contribution du canal vocal. Sur la figure (3.6.), on voit différentes fenêtres temporelles et leurs transformées de Fourier. On peut remarquer qu'une fonction limitée dans le domaine temporel n'est pas en général, limitée dans le domaine fréquentiel et inversement.
Une dernière remarque est relative au spectre de puissance. (transformée

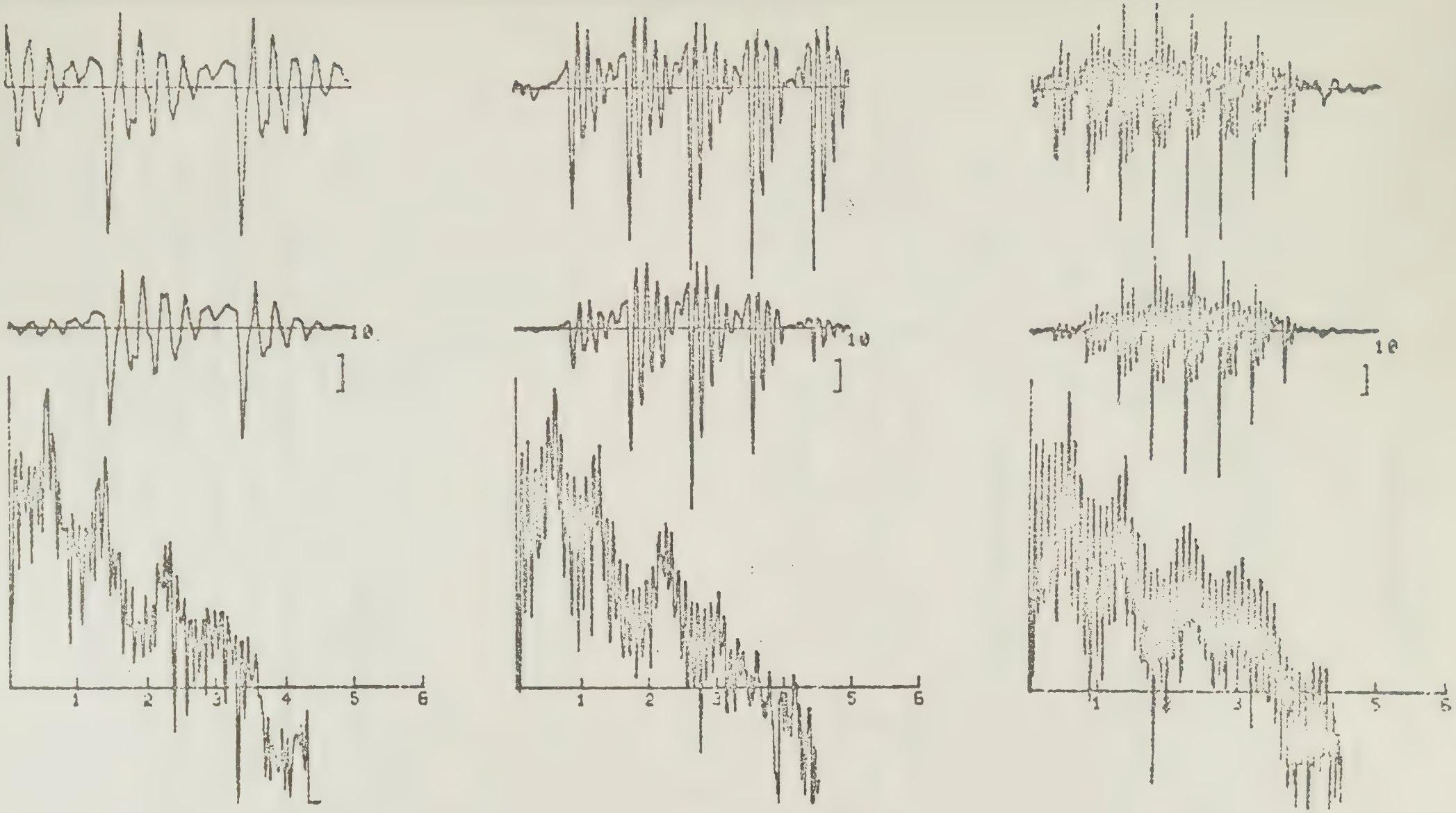
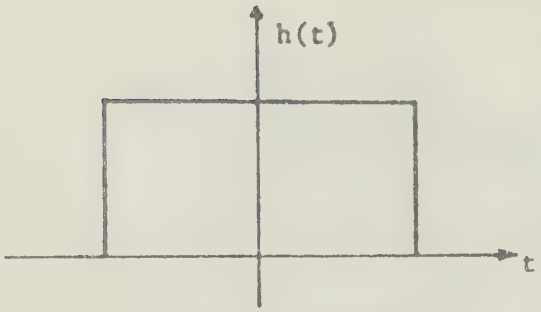
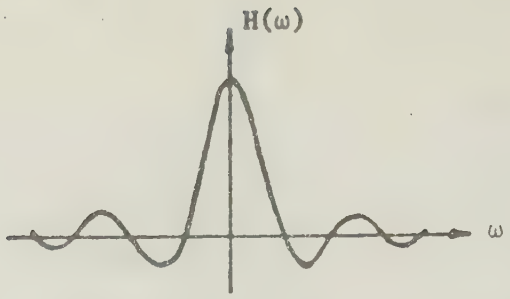


Figure 3.5 : Spectres du signal vocal calculés sous une fenêtre de Hamming de durée différente.

FENETRE RECTANGULAIRE

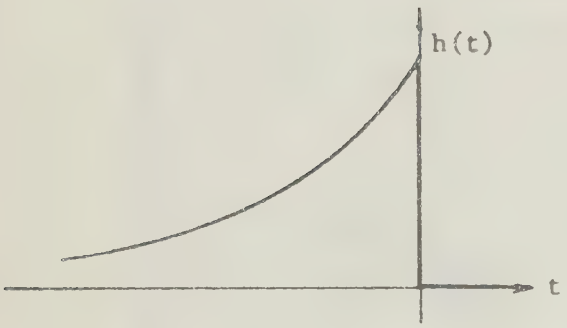


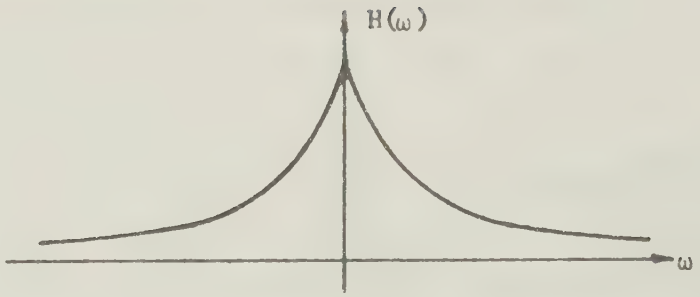


$h(t) = 1$ pour $|t|<T$
0 ailleur

$$H(\omega) = 2T\frac{\sin\omega T}{\omega T}$$

FENETRE EXPONENTIELLE

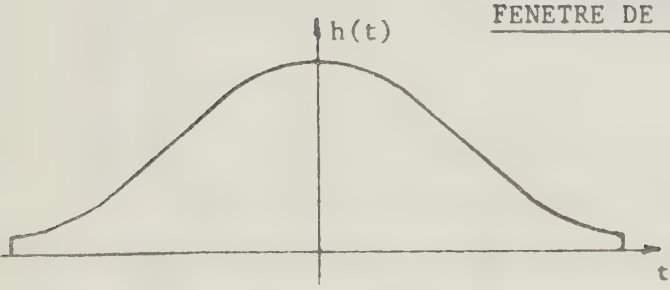


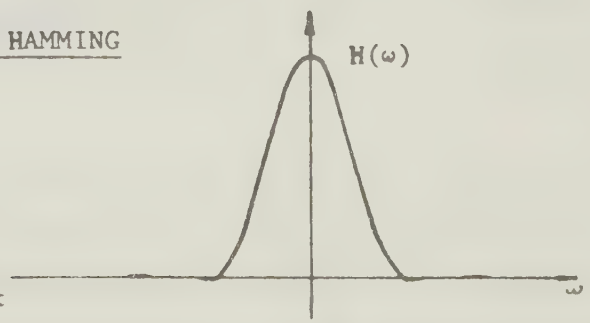


$h(t) = e^{at}$ pour $t<0$
0 ailleur

$$|H(\omega)| = \frac{1}{a^2+\omega^2}$$

FENETRE DE HAMMING





$$h(t) = .54 + .46\cos(\pi t/T)$$

$$H(\omega) = 1.08T\frac{\sin(\omega T)}{T} + .46T\frac{\sin(\omega+\frac{\pi}{T})T}{(\omega+\frac{\pi}{T})T} + .46T\frac{\sin(\omega-\frac{\pi}{T})T}{(\omega-\frac{\pi}{T})T}$$

Figure 3.6. : Différentes fenêtres temporelles et leurs spectres.

de Fourier de la fonction d' autocorrelation).

On aura ainsi pour la fonction d'autocorrelation :

$$v_k = \frac{1}{N} \sum_{i=0}^{N-1} s_i \, s_{i+k} \qquad \text{avec} \quad k = -(N-1), \ldots, 0, \ldots, (N+1)$$

Le spectre de puissance est alors :

$$P_k = \sum_{i=-(N-1)}^{N-1} v_i \exp(-j \, 2\pi k \delta f i T) \qquad k = 0, \ldots, 2N-1$$

Il est à remarquer que le pas de calcul δf est ici la moitié de celui de la formule (3.15) . La fonction d'autocorrelation , est en effet , définie pour $i = (-N+1), \ldots, 0, \ldots, (N-1)$ soit sur 2N points (au lieu de N pour le signal) le pas fréquenciel est donc :

$$\delta f = \frac{1}{2NT}$$

3.3.2.2. Recherche de l'information phonèmique dans le spectre à court terme . Déconvolution du signal vocal . Cepstre .Enveloppe

Les variations du canal vocal au cours de la phonation n'étant pas très rapides , un spectre à court terme calculé sous une fenêtre glissante , avec un pas donné , tout au long d'une phrase pourrait suivre ces variations et constituer un moyen de calcul très précieux . En observant les spectres donnés précédemment (Figure 3.5.) on peut remarquer des pics dont la fréquence est très caractéristique du phonème prononcé . Ces pics correspondent aux fréquences naturelles de résonance du canal vocal .Pourtant un tel spectre est comme on l'a remarqué très "modulé " par l'influence de la source excitatrice du canal vocal soit si elle est périodique soit s'il s'agit d'un bruit

Ceci rend la détection des pics correspondant aux fréquences de résonance (formants) très difficile voire impossible, en automatique . Il convient donc de séparer les deux influences (source -canal vocal) cette séparation est relativement logique vu que les deux processus sont relativement indépendants .

Le modèle simple de production de la figure (3.7) nous aidera à cette séparation .

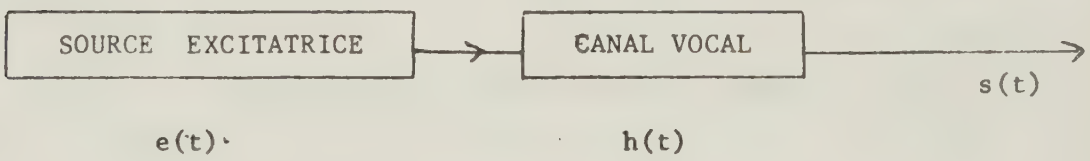


Figure 3.7.

L'effet du canal vocal peut être complètement déterminé en fonction de sa réponse impulsionnelle h(t) , de façon à considérer la sortie (signal vocal) comme la convolution de e(t) et de h(t)

$$s(t) = e(t) * h(t) \qquad (3.17.)$$

La relation (3.17.) dans le domaine fréquentiel pourrait s'écrire :

$$S(\omega) = E(\omega) . H(\omega) \qquad (3.18.)$$

où les majuscules désignent les transformées de Fourier des fonctions initiales .

La source excitatrice est quasipériodique ainsi que le signal vocal .

Cette périodicité doit se trouver sur le spectre . Le spectre de puissance contient donc des harmoniques espacés de T_o^{-1} HZ (T_o : période fondamentale). alors il est " périodique " avec une période égale à l'inverse de la période du signal initial . Pour mesurer cette période il suffirait , par conséquent , de prendre la transformée de Fourier du spectre de puissance . Cette transformée présenterait un pic caractéristique de cette période. La transformée de Fourier du spectre de puissance n'est rien d'autre que la fonction d'autocorrélation du signal initial.

$$v(\tau) = \mathcal{F}(|S(\omega)|^2) = \mathcal{F}(|E(\omega)|^2 . |H(\omega)|^2)$$

$$v(\tau) = v_e(\tau) * v_h(\tau)$$

Les effets donc du canal vocal et de la source vocale sont convolués sur les fonctions d'autocorrelation. Ceci donne un résultat qui n'est pas satisfaisant pour la mesure du pitch . (figure 3.8.)
Pourtant nous savons qu'il y a une fonction , le logarithme , qui transforme un produit en une somme . La relation (3.18.) s'écrit :

$$\text{Log}(|S(\omega)|^2) = \text{Log}(|E(\omega)|^2 . |H(\omega)|^2) =$$

$$= \text{Log}(|S(\omega)|^2) + \text{Log}(|H(\omega)|^2) \qquad (3.19)$$

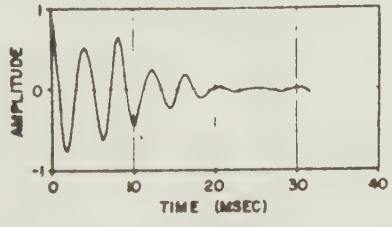
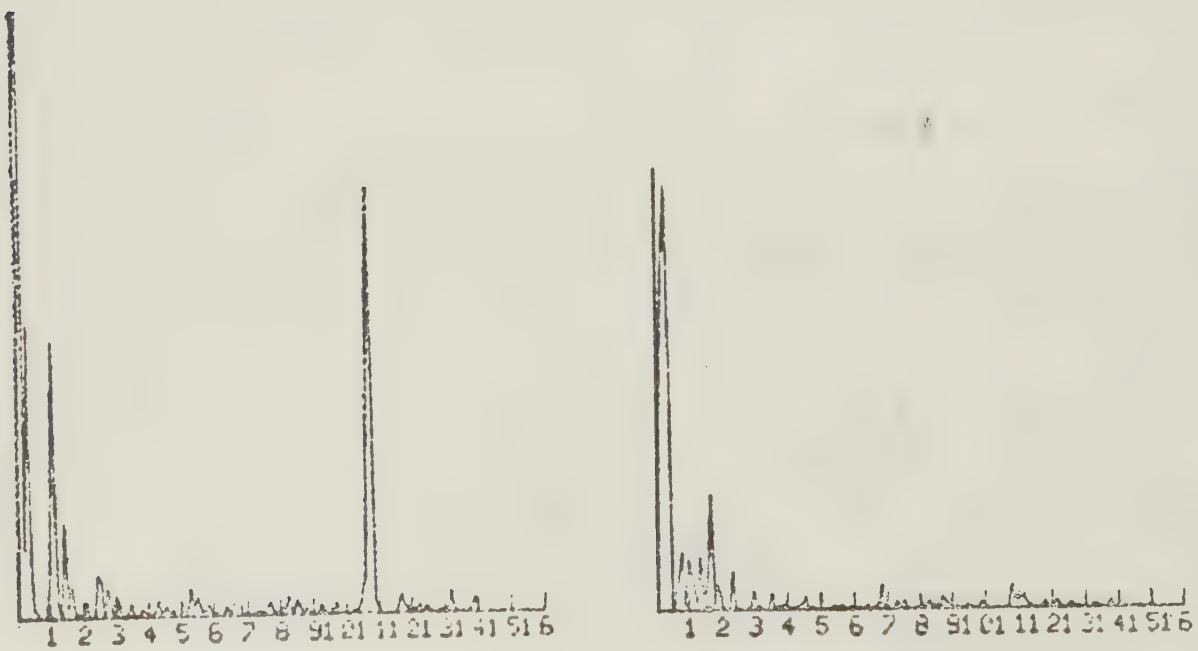


Figure 3.8. : Fonction d'autocorrelation normalisée (voyelle i) . (figure tirée de/14/)

et en prenant la transformée de Fourier :

$$\mathcal{F}(\text{Log } S(\omega)^2) = \mathcal{F}(\text{Log } E(\omega)^2) + \mathcal{F}(\text{Log } H(\omega)^2) \qquad (3.20)$$

Les deux effets sont maintenant plutôt additifs que convolués .
NOLL /4/, /13/ défini sous le nom "cepstre" le carré de la transformée de Fourier du logarithme du spectre de puissance . D'autres ensuite ont donné des définitions légérement différentes . Ainsi Oppenhein définit le cepstre comme étant la transformée de Fourier inverse du logarithme de la transformée de Fourier quand l'information sur la phase n'est pas retenue et le cepstre complexe comme étant la même transformée mais quand l'information sur la phase est retenue .
Sur la figure (3.9.) on voit un pic caractéristique de la fréquence fondamentale sur un cepstre calculé avec la définition de NØLL . Ce pic n'existe pas pour un son sourd . (figure 3.9.)
La détection de ce pic constitue une très bonne méthode de mesure du pitch . Pourtant il faut noter qu'il s'agit d'un pitch moyen sous la fenêtre considérée ; mais étant donné que le pitch varie très peu surtout sur le même phonème , l'estimation donnée avec cette méthode est assez précise .



Son sonore

Son sourd

Figure 3.9. : Cepstres calculés d'après la définition de NOLL .

NØLL a élaboré un algorithme de détection de ce pic . Cet algorithme sert en même temps comme algorithme de décision son voisé - non-voisé . Une mesure automatique de la fréquence fondamentale peut donc s'effectuer sur une phrase donnée . La figure (3.10.) explique l'automatisation du procédé .

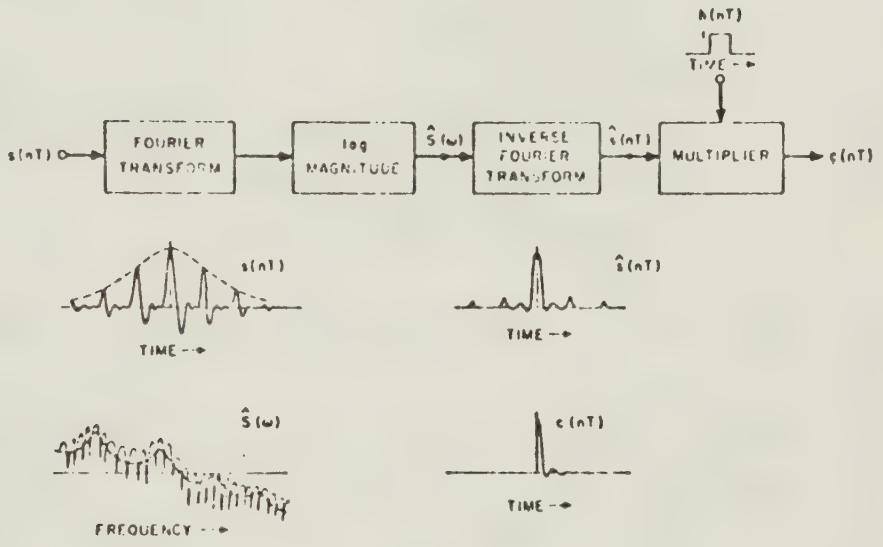


Figure 3.10. Automatisation du procédé de mesure du pitch sur une phrase .
(figure tirée de /60/)

Pourtant il faut noter que le niveau dynamique du pic en question est très variable . Ceci est surtout dû au fait que les différents segments sont disposés chaque fois d'une manière variable sous la fenêtre et que le nombre de ces segments qui n'est pas toujours le même , ce qui fait augmenter ou diminuer chaque fois la contribution du pitch . En outre la composante à l'origine du cepstre ne peut pas être utilisée comme base de normalisation . Cette composante , étant très grande , est dûe au fait que le logarithme ne change pas en général de signe sur le spectre , elle est donc très variable dependant de la forme du spectre . Ces raisons sont à l'origine de certaines difficultés de l'algorithme au niveau décision "voisé - non voisé" . D'autres difficultés viennent d'un phénomène qui s'appelle dédoublement du pitch (doubling) . Ce phénomène donne naissance à un pic très prononcé correspondant à une fréquence égale à celle du second harmonique (voir figure 3.11.) .

ceci peut entrainer de fausses mesures du pitch .

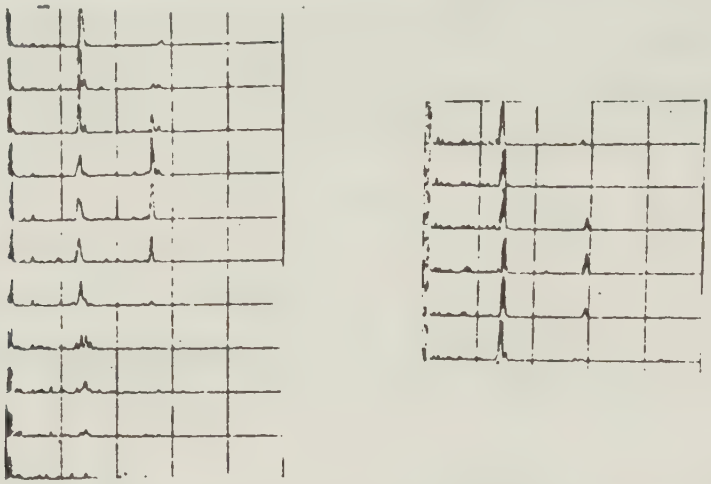


Figure 3.11. : Le phénomène de dédoublement du pitch.

Ces difficultés et d'autres encore empêchent l'algorithme de pouvoir constituer un outil infaillible de détection du fondamental
Il ne faut pas terminer sans noter deux grands avantages de la méthode
Ces avantages sont les suivants :

1. Elle est résistante à des bruits additifs
2. Elle est résistante à des distortions des amplitudes .

Nous avons vu que les contributions de l'excitation et du canal vocal se trouvent séparées sur le cepstre . L'information concernant l'excitation se trouve à des multiples de la période du pitch , tandis que l'information concernant le canal vocal se trouve localisée près de l'origine. Si on utilise cette information comme entrée pour une nouvelle transformée de Fourier on doit trouver à la sortie , le spectre lissé , c'est à dire le spectre complétement dépourvu de la variation rapide . Il s'agit finalement de l'enveloppe du spectre initial . Ici les pics correspondant aux fréquences de résonance du canal vocal sont mis en evidence et peuvent facilement être détectés avec un simple algorithme de détection (pick - picking algorithms) / 3 / . Ce genre d'algorihmes peuvent donner des appareils simples de détection câblés en hard (pick - pickers) / 3 /
Sur la figure (3.12.) on voit l'enveloppe du spectre de la figure (3.5.)

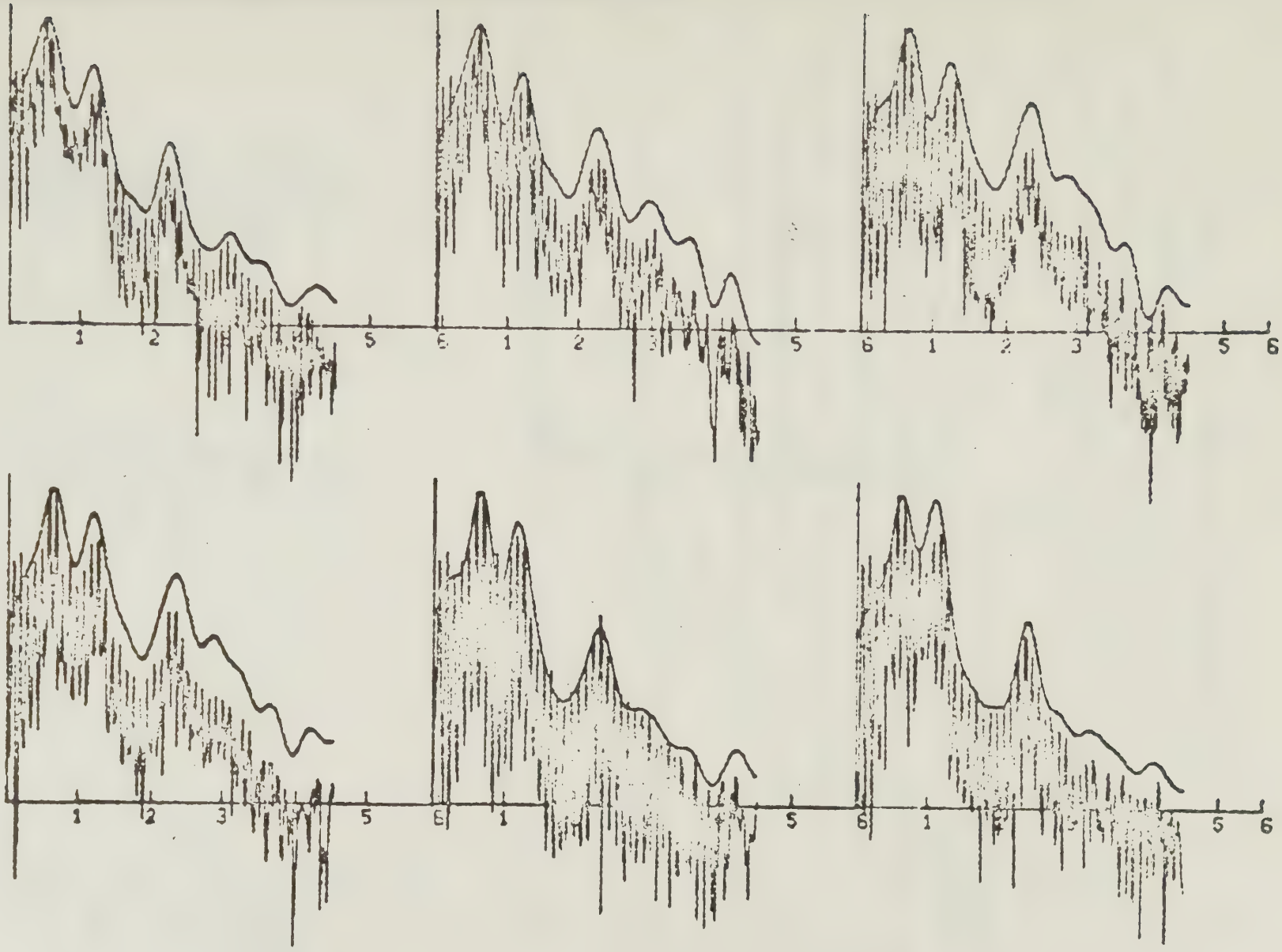


Figure 3.12 : Evolution de l'enveloppe du spectre sur un phonème /a/.

L'évolution de cette enveloppe sur un phonème /a/est donnée .
Sur la figure (3.13.) on voit les opérations qui conduisent à ce résultat La même procédure est valable pour la déconvolution des sons sourds , la partie du cepstre près de l'origine constituant la partie caractéristique de l'enveloppe ; le reste étant en relation plutôt avec l'entrée (bruit dans ce cas).

Considérations pratiques :1. Le choix de la largeur de la fenêtre temporelle est très délicat . En effet pour pouvoir détecter facilement le pitch il faut une fenêtre assez large qui comprendrait plusieurs périodes de pitch . En outre pour pouvoir détecter des formants une fenêtre assez étroite doit s'appliquer surtout sur les endroits des transitions où les formants varient beaucoup . Finalement un compromis doit être choisi. Ainsi une fenêtre comprenant (3÷4)périodes donne des résultats assez satisfaisants .

2. Grâce aux symétries de la transformée de Fourier rapide , on peut calculer deux transformée réelles par un seul calcul.

3. La FFT peut trouver des applications pour le calcul des convolutions et autocorrelations plus rapidement .

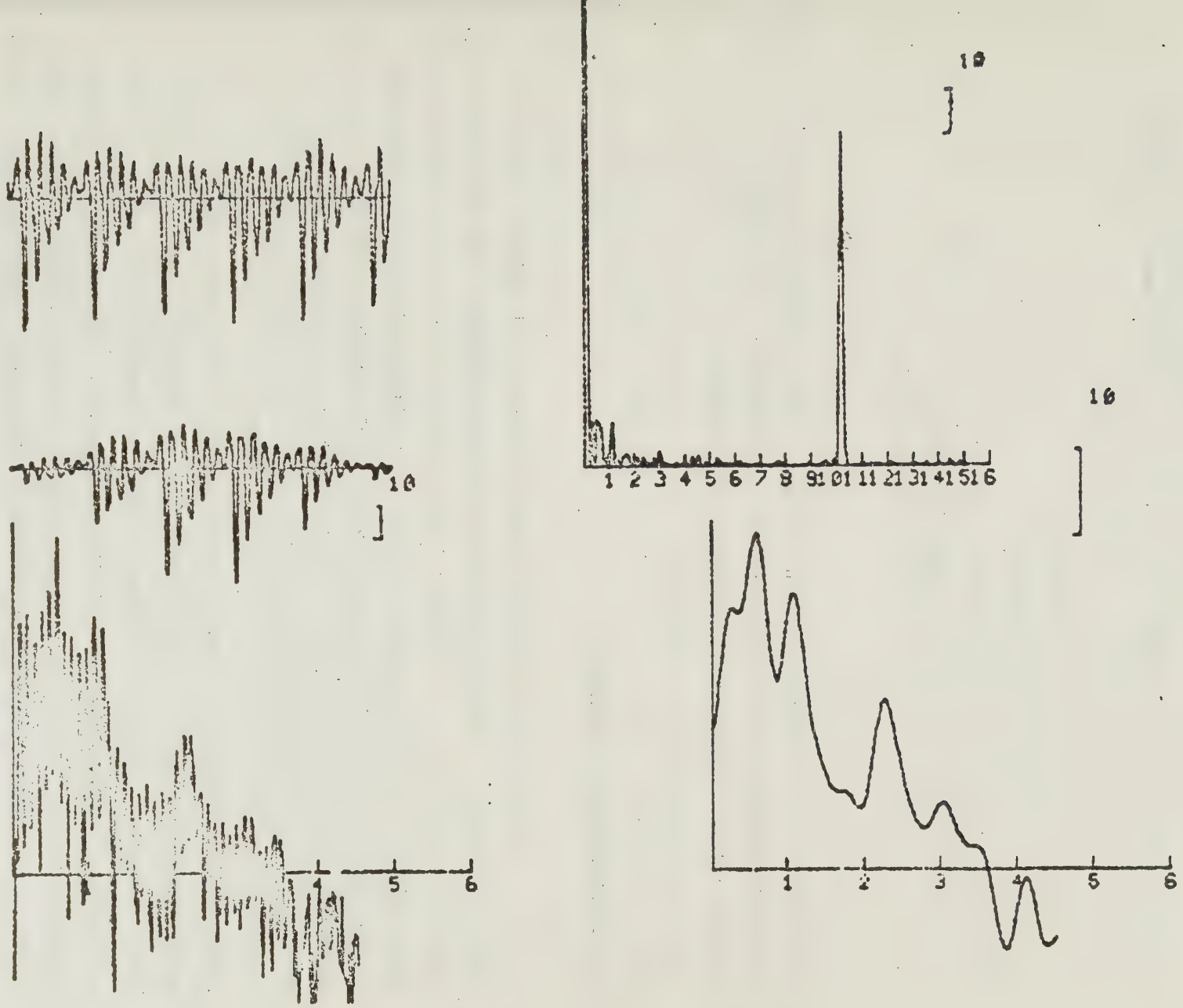


Figure 3.13. : Stades intermédiaires qui conduisent à l'enveloppe du spectre.

3.3.2.3. Applications

L'analyse de la parole à l'aide du spectre logarithmique et du cepstre peut trouver plusieurs applications dans les domaines suivants :

. Détection du fondamental
. Détection des formants
. Synthèse de la parole
. Transmission économique des données
. Reconnaissance automatique des locuteurs

Nous avons vu la détection du fondamental ; Nous examinerons par la suite encore trois applications , la détection des formants , la synthèse de la parole , et la transmission économique des données .

A. Détection des formants.

Le calcul de l'enveloppe étant accompli, il suffirait pour calculer les formants d'avoir un simple algorithme de détection. Pourtant le problème de détection des formants n'est pas aussi facile. Les raisons sont les suivantes : 1) Parfois des pics parasites apparaissent sur l'enveloppe. Il faut alors pouvoir reconnaître les pics intéressants parmi ces pics parasites. 2) Les régions de variation des différents formants se chevauchent (fig. 3.14), ainsi parfois quand les formants sont très près, un seul pic apparaît au lieu de deux pics SCHAFFER - RABINER /16/ ont élaboré un algorithme de détection automatique où ils ont inclus des considérations de continuité de la variation des formants et de dépendance des valeurs calculées à l'instant courant de celles calculées auparavant. En outre, ils ont inclus dans le même algorithme des considérations traduisant les relations existant entre la fréquence et l'amplitude relative à chaque formant, ainsi qu'entre les amplitudes de deux formants consécutifs. Mais tout ceci n'a pas été suffisant vu le chevauchement des régions pour différents formants.

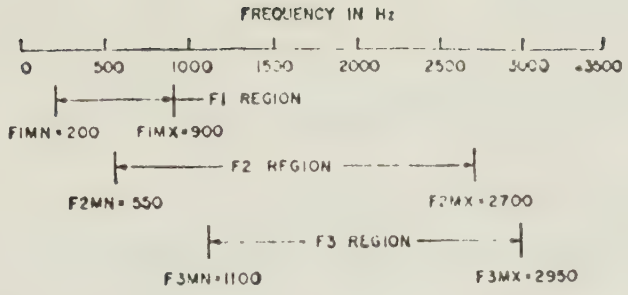


Figure 3.14. Régions de variation des trois premiers formants pour un locuteur masculin d'après SCHAFER et RABINER .

Ainsi dans le cas où un seul pic apparaît pour les deux formants il faudrait une analyse plus fine qui permettrait de séparer les deux pics . La première idée est de calculer non plus la transformée de Fourier mais la transformée en z sur un contour plus général qui passe plus près des pôles . Un algorithme rapide a été élaboré par SCHAFER et RABINER / 16/ qui effectue cette analyse . La seconde idée est d'utiliser un modèle pour le spectre /17/. On verra certains aspects de ces algorithmes dans le prochain chapitre .

A. 1 Chirp - z - Transform .

Nous avons vu les limites de la transformée de Fourier rapide quand elle est utilisée pour le calcul de l'enveloppe du spectre . En effet pour des formants relativement rapprochés (moins de 300 Hz) on n'arrive pas à distinguer deux pics séparés . Un seul pic apparaît donc pour les deux formants .

Nous savons que la transformée de Fourier discrète correspond au calcul de la transformée en z sur la circonférence du cercle unitaire dans le plan des z (ce qui correspond à l'axe imaginaire du plan de Laplace) Mais le calcul de la transformée en z serait possible sur n'importe quel autre contour . En particulier si le contour passe plus près des pôles que le cercle unitaire , il pourrait donner des pics beaucoup plus prononcés . La transformée en z " à court terme " est définie comme suit :

$$X_k = X(z_k) = \sum_{n=1}^{N-1} x_n \, z_k^{-n}$$

Posons $z_k = A\, W^{-k}$ $\qquad k = 0, 1, \ldots, (M-1)$

$$X_n = \sum_{n=0}^{N-1} x_n \, A^{-n} \, W^{nk} \qquad k = 0, \ldots, (M-1) \qquad (3.21)$$

où M est un entier choisi arbitrairement et A et W sont des nombres de la forme :

$$A = A_o \exp(j\, 2\pi\theta_o) \qquad (3.22)$$

$$W = W_o \exp(j\, 2\pi\phi_o) \qquad (3.23)$$

Le cas A=1 , N=M , W= exp(-j2π/N) correspond au calcul de la FFT classique .

$$X_k = \sum_{n=0}^{N-1} x_n \exp(-j\, 2\pi\, nk/N) \quad , k = 0, 1, \ldots, (N-1)$$

Sur le plan des z le contour commence au point z = A et va vers l'intérieur ou l'extérieur du cercle suivant la valeur de W . Sur la figure (3.15) on voit un tel contour , le contour spiral sur le plan des z correspond à une droite sur le plan de Laplace.

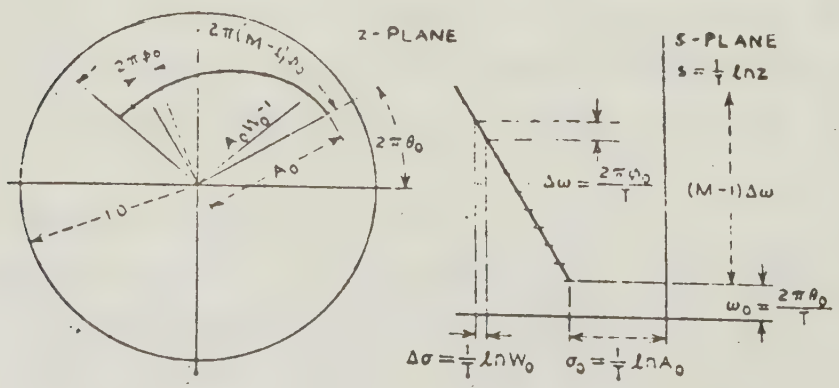


Figure (3.15.) :

La transformée en z est calculée sur un contour plus général comme on le voit sur la figure . Le contour correspondant au plan de Laplace est donné à droite (figure tirée de /16/).

En effet le point général du contour dans le plan des z est :

$$S_k = S_o + k(\Delta\sigma + j\Delta\omega) = (\ln A - k \ln W)/T$$

$$k = 0,1,\ldots,(M-1)$$

Vu que A et W sont des nombres complexes choisis arbitrairement , nous pouvons voir que les points S_k se trouvent sur un segment de droite

Sur la figure (3.16.) on voit l'exemple de 5 spectres calculés par RABINER pour le phonème /i/ dans le mot /beet/ sur les 5 contours illustrés à gauche .

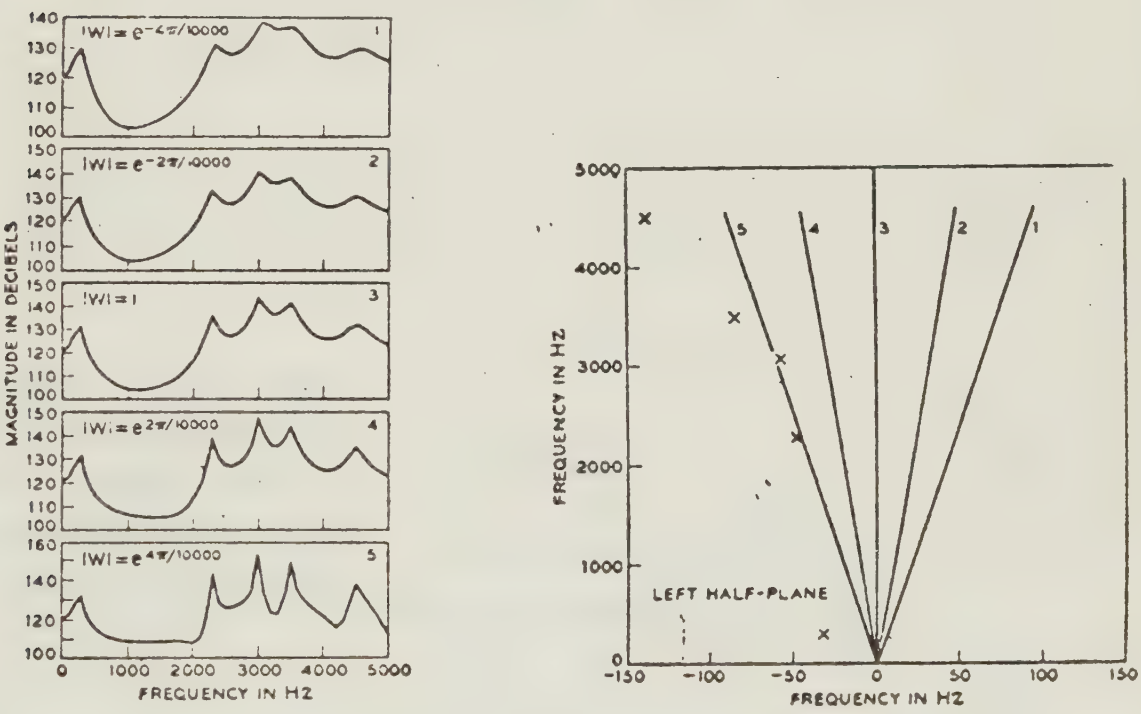


Figure 3.16 . Cinq spectres différents calculés sur les cinq contours de la figure d'après Rabiner /16/.

- Un algorithme de calcul rapide de la transformée en z sur un contour spiral du plan en z a été élaboré par SCHAFER et RABINER . Cet algorithme est basé sur la " F.F.T " et a été publié sous le nom de " Chirp z transform

(CZT) (par analogie à la FFT). /16/.
Le calcul de la transformée en z basé sur l'équation (3.21) demande un temps assez important étant donné qu'il faut effectuer (NM) multiplications complexes en total. Un certain nombre de transformations sont à l'origine d'une réduction considérable du nombre d'opérations à effectuer.

La "CZT" appliquée directement sur le cepstre peut donner une séparation de formants autrement confondus (voir figure 3.17).

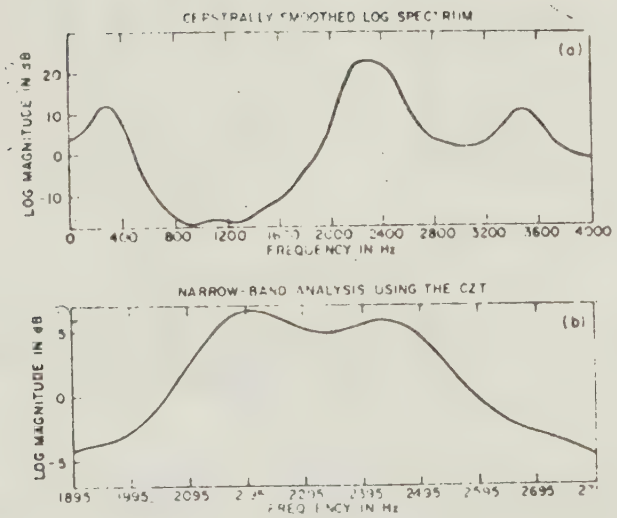


Figure 3.17. : Séparation de deux pics du 2è et 3è formant autrement confondus , en appliquant la CZT directement sur le cepstre. D'après RABINER /16/

A2) Détection des formants en utilisant un modèle du spectre.

- Avec cette méthode , un modèle du spectre est comparé avec un spectre mesuré . Le spectre mesuré est l'enveloppe du spectre calculé avec la méthode de déconvolution du signal vocal . La différence entre le spectre théorique,dérivé du modèle,et le spectre mesuré est considérée par la suite Un critère des moindres carrés est utilisé . Pour le modèle une série de résonateurs est considérée . Un spectre à court terme peut s'exprimer comme étant :

$$\hat{S}(f) = \prod_{i=1}^{K} AT_i(f,f_i)\ G(f)\ R(f) \qquad (3.24)$$

où f est la fréquence , A est l'amplitude et S(f) est le spectre, G(f) et R(f) sont les spectres de la source glottale et de radiation buccale respectivement . Une approximation pour G(f) et R(f) est adoptée:

$$G(f)\ R(f) = \frac{f/100}{1 + (f/100)^2} \qquad (3.25)$$

$T_i(f,f_i)$ sont des résonateurs définis par :

$$T_i(f,f_i) = \frac{(jf_i + a_i)\ (-jf_i + a_i)}{(jf-jf_i-a_i)\ (jf + jf_i-a_i)} \qquad (3.26)$$

où f_i sont les fréquences de résonance et a_i les facteurs d'amortissement Dans la pratique K est pris égal à 3 . Les facteurs d'amortissement sont pris constants .

Soit S(f) le spectre mesuré . L'erreur est définie comme étant :

$$Q = \int (\hat{S}(f) - S(f))^2 \, df \qquad (3.27)$$

Un système de trois équations est résolu:

$$\frac{\partial}{\partial f_i} \left(\int (\hat{S}(f) - S(f))^2 \, df \right) = 0 , \quad i = 1,2,3 \qquad (3.28)$$

Il s'agit d'un système non linéaire et une technique de Newton-Raphson est utilisée pour la résolution. Après développement de (3.28) en série de Taylor, en posant $f_i = f_i^o + \delta f_i$, seuls les termes jusqu'au second ordre $(\delta f_i)^2$ sont retenus).

Cette méthode a été appliquée par OLIVE /17/. Les formants détectés ont été utilisés pour la synthèse vocale.

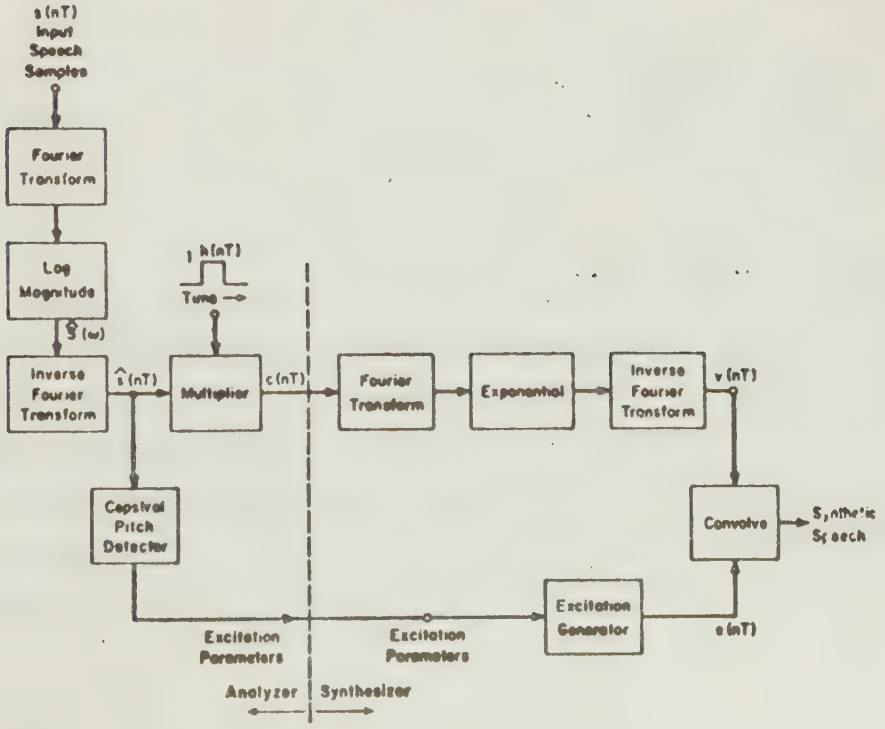
B. Synthèse de la parole.

Deux systèmes tout à fait différents dans leur conception ont été basés sur cette technique d'analyse de la parole .

B1) Synthétiseur d'Oppenheim . (Homomorphic vocoder)

Quelques échantillons se situant près de l'origine du cepstre sont envoyés au synthetiseur. Des paramètres supplémentaires comme :
1) valeur du pitch
2) puissance du signal
3) un paramètre binaire caractérisant le voisement, sont envoyés aussi au synthétiseur .

Ces informations conformément à l'analyse faite précedemment doivent être suffisantes pour la reconstitution du signal .
Ce synthétiseur a la configuration suivante : (voir Fig 3.18)



Sur la même figure on voit l'analyseur attaché à ce synthétiseur. Les quelques échantillons du début du cepstre (portant toute l'information sur le canal vocal) résultant de l'analyse sont transformés en une réponse impulsionnelle , qui convoluée avec une fonction d'excitation produit la parole synthétique .

La fonction d'excitation est générée sur la base des données relatives aux variations du pitch . L'existence du fondamental est d'autre part spécifiée avec un paramètre binaire . Durant le voisement l'excitation consiste en une série d'impulsions espacées de la valeur de la période calculée . Dans le cas contraire une série de nombres aléatoires est utilisée comme excitation . Les amplitudes sont ajustées pour avoir l'énergie désirée à la sortie .

La réponse impulsionnelle est obtenue par le cepstre en calculant la transformée de Fourier des quelques valeurs du cepstre , suivie d'une transformation exponentielle et d'une transformée de Fourier inverse. Si C_n est le n^{e} échantillon(ou canal)du cepstre on aura :

$$\hat{v}(nT) = \hat{v}(-nT) = C_n \qquad 0 < n < M-1$$

$$\hat{v}(nT) = \hat{v}(-nT) = 0 \qquad M < n \qquad (3.29)$$

où $\hat{v}(nT)$ est l'entrée à la transformation suivante . De cette façon le cepstre est pris comme une fonction paire et la réponse impulsionnelle résultante sera aussi une fonction paire . Cette réponse impulsionnelle aura donc une phase zéro . (elle a une transformée de Fourier réelle). Il y a un moyen de transformer le cepstre dans le but d'introduire des caractéristiques de phase $\neq o$, sans changer l'amplitude du spectre . Pour faire ceci Oppenheim a écrit le cepstre de la façon suivante :

$$\hat{v}(nT) = \begin{cases} 2C_n & 0 < n < M-1 \\ C_n & n=0 \\ 0 & n < 0 \end{cases} \qquad (3.30)$$

On peut facilement vérifier que la réponse impulsionnelle obtenue avec la défintion (3 . 30.) n'a pas de phase zero puisque sa transformée de Fourier a une partie imaginaire différente de zero tout en gardant la même amplitude .
La réponse impulsionnelle ainsi produite est une réponse impulsionnelle de phase minimale . Comme l'a constaté OPPENHEIM, l'introduction d'une phase minimale dans la parole synthètique lui donne des caractéristiques d'une voix naturelle . Le nombre de canaux du cepstre utilisé initialement était de l'ordre de 32 pour être réduit par la suite à 26 (/60/, /61/).
Les 26 canaux ont été quantifiés à 6 bits / canal ce qui correspond à un taux de transmission de 7800 bits /sec) .
Il faut noter que la conception de ce système ressemble beaucoup à celle d'un vocoder à canaux .

B2) Synthétiseur des Schafer et Rabiner.

Le calcul au niveau analyseur ne s'arrête pas au cepstre comme précedemment , mais il continue jusqu'à l'enveloppe . La détection des formants est effectuée et ce sont les formants qui sont envoyés au synthetiseur .
Le synthétiseur pour les sons voisés a la configuration de la figure 3.19 .

La boite notée générateur d'impulsions produit une série d'échantillons unitaires espacés de T_o ; T_o étant la période du pitch qui est fonction du temps $T_o = T_o(t)$
Le multiplieur A est destiné à contrôler le gain de façon à avoir l'intensit désirée à la sortie .

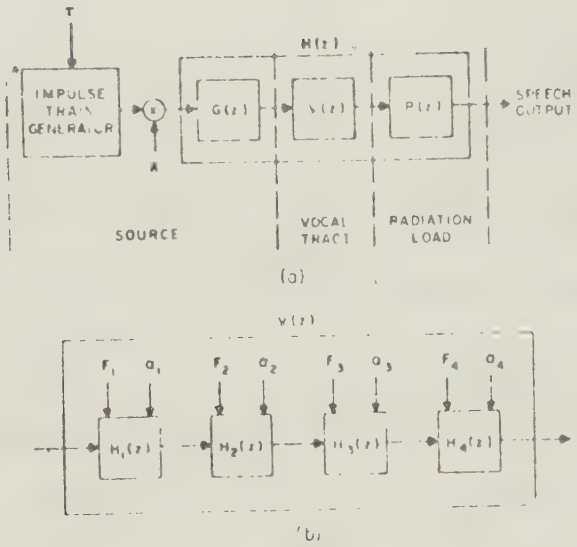


Figure 3.19. Configuration du Synthètiseur de Schaffer et Rabiner pour les voyelles .

La cascade des boites : $G(z)\ R(z)$ est une approximation de la combinaison du spectre de la source glottale et du spectre de radiation.

On aura :
$$G(z)\ R(z) = \frac{(1-\exp(-aT)).(1 + \exp(-bT))}{(1-z^{-1}\exp(-aT)).(1 + z^{-1}\exp(-bT))}$$

où a et b sont des constantes qui caractérisent le locuteur (et probablement la phrase). Des valeurs représentatives de a et b données par Rabiner sont $a = 400\ \pi$, $b = 5000\ \pi$. Pour ces valeurs de a et b G (z) R (z) évalué sur le cercle unitaire donne la variation suivante: (voir figure 3.20.)

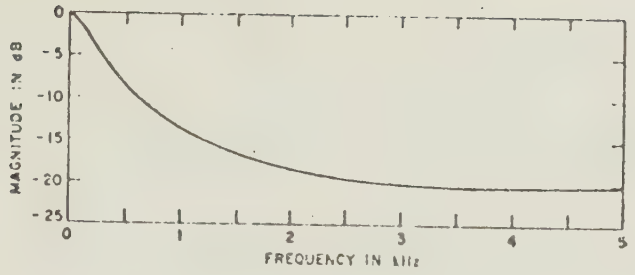


Figure 3.20. Variation de G(z) R(z).

Le système V (z) est un système discret variable avec le temps qui consiste en une connexion en cascade des résonateurs comme on voit sur la figure 3.19.b.

$$V(z) = \prod_{k=1}^{4} H_k(z) \qquad \text{où} \qquad H_k(z) = \frac{(1-2\exp(-a_k T)\cos(\omega_k T) + \exp(-2a_k T))}{(1-2z^{-1}\exp(-a_k T)\cos(\omega_k T) + z^{-2}\exp(-2a_k T))}$$

où $\omega_k = 2\pi F_k$; T= fréquence d'échantillonage.

Le but de ce système est d'approximer les caractéristiques de transmission du canal vocal . Les fréquences des formants F_k et les bandes passantes $2a_k$ sont variables avec le temps et simulent le processus de phonation. Un pôle et un zéro sont encore nécessaires pour représenter les nasales. Les premières expériences entreprises par Schaffer et Rabiner ont démontré que la qualité de la voix synthetique produite n'est pas forcement altérée si ce zéro et ce pôle ne sont pas présents .

Pour les sons sourds un modèle différent est utilisé , pour approximer les caractéristiques du canal vocal . Ces caractéristiques sont résumées par la fonction de transfert qui est :

$$H(z) = \frac{A.B}{C.D}$$

où $A = 1-2\exp(-bT)\cos(2\pi F_p T) + \exp(-2bT)$

$B = 1-2\exp(-bT)\cos(2\pi F_z T)z^{-1} + z^{-2}\exp(-2bT)$

$C = 1-2\exp(-bT)\cos(2\pi F_z T) + \exp(-2bT)$

$D = 1-2\exp(-bT)\cos(2\pi F_p T)z^{-1} + z^{-2}\exp(-2bT)$

Le circuit de compensation de l'influence de la source glottale et de radiation reste le même.

Sur la figure (3.21) on donne la configuration globale du synthétiseur. On remarque que les productions des sons sonores et des sons sourds restent complètement indépendantes.
La figure (3.22) nous montre les données utilisées par ce synthétiseur, c'est-à-dire la variation des trois formants ainsi que la variation du pitch tout au long d'une phrase.

L'analyseur attaché à ce synthétiseur est un système asynchrone de détermination du pitch, du gain et des trois premiers formants. La détermination des formants se fait, après détection, sur l'enveloppe calculée par la méthode du cepstre. La période fondamentale est calculée en même temps. Vu que le système est asynchrone, c'est une mesure du pitch qui est effectuée et non pas la détection de l'arrivée des impulsions individuelles. L'avantage de ce traitement asynchrone est qu'on évite de déterminer les arrivées des impulsions du pitch ce qui est un problème très difficile. L'inconvénient est que les formants et la période du pitch sont moyennés sous la fenêtre temporelle. Ceci d'après les auteurs n'a pas provoqué d'altérations graves du signal acoustique synthétique sur les transitions formantiques et sur les brusques changements des périodes du pitch au moins sur les voix masculines. Les paramètres suivants sont maintenus constants pendant tout le processus de phonation :

$a_1 = 60\pi$, $a_2 = 100\pi$, $a_3 = 120\pi$, $a_4 = 175\pi$, $a = 400\pi$, $b = 5000\pi$

$F_4 = 4000$ Hz .

Un taux de transmission de 1000 bits / sec a été obtenu .

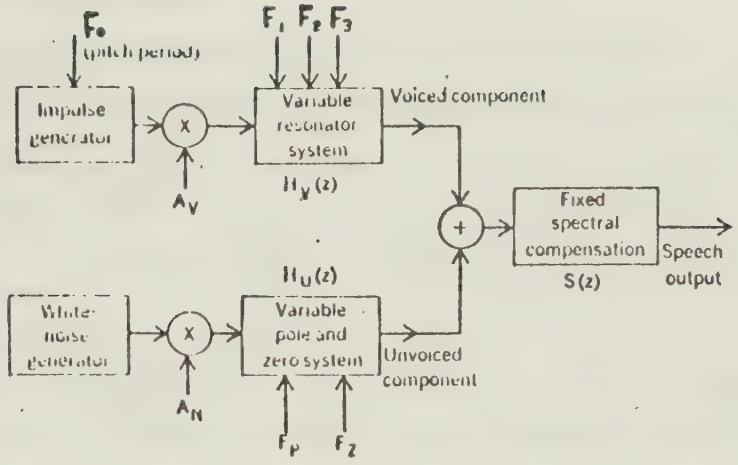


Figure 3.21. : Configuration du synthètiseur B2

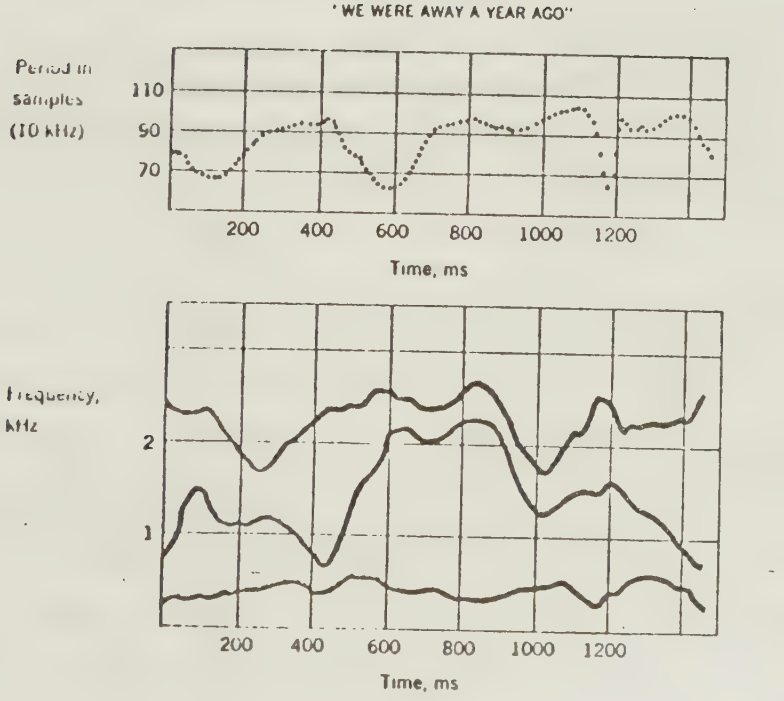


Figure 3.22. : Informations utilisées par le synthètiseur B2.

3.3.3 Méthodes temporelles

3.3.3.1. Méthode des passages à zéro.

Il s'agit d'une méthode de paramètrisation rapide du signal vocal . Deux sortes de paramètres différents sont en général utilisés .

. Le nombre de passages à zéro sur un segment de durée spécifiée.
. Les propriétés statistiques de dispersion des intervalles entre deux passages à zéro sur un segment de durée spécifiée.

Certaines relations simples qui existent entre le spectre de puissance et le taux des passages à zéro moyen d'un processus ergodique , constituent le fondement théorique de la méthode / 18 / . Si z est une estimation du taux des passages à zéro d'un signal s(t) et P(f) le spectre de puissance du signal en question (voir § 3.3.2.1.) nous aurons la relation :

$$z_o = k_o \left[\int_{-\infty}^{+\infty} f^2 \, P(f) \, df \Big/ \int_{-\infty}^{+\infty} P(f) \, df \right]^{1/2}$$

où k_o est une constante
Cette relation est valable d'une manière plus générale sur les dérivées d'ordre n du signal . Si z_n est le nombre de passages à zéro de $(d^n s(t) / dt^n)$ on aura :

$$z_n = k_n \left[\int_{o}^{\infty} f^{2n+2} \, P(f) \, df \Big/ \int_{o}^{\infty} f^{2n} \, P(f) \, df \right]^{1/2}$$

Pour un processus gaussien : $k_n = 2$ (voir /18/).

Si : $x(t) = A \sin (2\pi a (t + \theta))$ où A est une constante et θ une variable aléatoire uniformément distribuée entre 0 et 1/a le calcul direct donne $z_n = 2a$. Ceci nous conduit à penser qu'une relation pourrait exister entre le nombre des passages à zéro et la fréquence

dominante d'un signal . L'approche cherchée est la suivante : Pouvoir calculer les formants du signal vocal en comptant les passages à zéro sur un intervalle donné .

Soit $$z_o = \sum_{k=1}^{N} \left[1 - \mathrm{Sgn}\ (s_{k+1})\ \mathrm{Sgn}\ (s_k) \right] \Big/ 2$$

où Sgn (X) = signe (X)

Pour les voyelles ITO et DONALDSØN /19/ ont démontré la relation suivante :

$$\left[\frac{z_o}{k_o} \right] = f_1^2 \left[1 + R_2^2\ (f_2 / f_1)^2 + R_3\ (f_3 / f_1)^2 \right] \Big/ (1 + R_2^2 + R_3^3)$$

où f_i : les fréquence des formants .

R_i : Des quantités constantes en relation avec les caractéristiques de chaque résonance .

Si le formant i est dominant R_i est beaucoup plus grand pour ce formant ($R_i \gg R_j$, $j \neq i$) et $R_i \gg 1$. Donc $(z_o / k_o) \simeq f_i$. Pour pouvoir calculer tous les formants , une procédure de filtrage passe bande doit être utilisée qui rendra chaque formant dominant dans une zone de fréquences . Pourtant les limites des fréquences des formants étant très variables et pas bien définies ce calcul ne peut être que très approximatif. Par conséquent cette méthode ne peut pas être utilisée avec succès pour l'estimation des formants .

Pourtant d'autres applications sont possibles : MAISSIS a utilisé les passages à zéro pour normaliser ses paramètres dans le cadre de la reconnaissance de la parole / 20 / (voir aussi CH4) DE MORI / 21 / et d'autres utilisent les distributions des intervalles entre deux passages à zéro pour créer un profil pour chaque phonème,

dans le cadre de la reconnaissance de la parole . Nous voyons les courbes de ces distributions (voir fig 3.23.) pour différents phonèmes dans la zone du premier et du second formant extraites de /21/ DE MORI a défini des groupes standardisés des intervalles des passages à zéro . Ainsi un intervalle de durée τ appartient au groupe i si

$$\tau_i < \tau \leqslant \tau_{i-1}$$

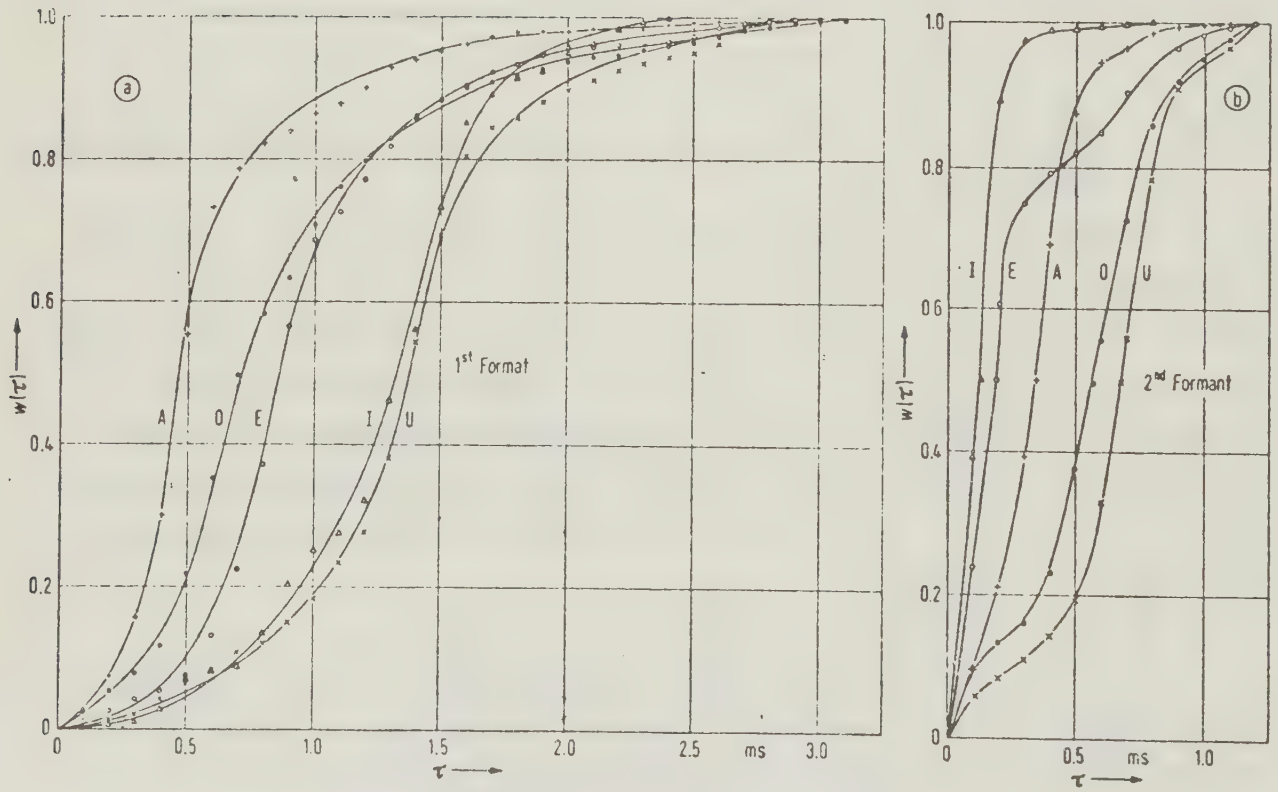


Figure 3.23. : Distribution statistique des intervalles entre deux passages à zéro à la sortie (a) d'un filtre passe bas , (b) d'un filtre passe haut .

Pour caractériser les phonèmes , il a utilisé les vecteurs suivants:

$$R_h(nT) = \left[R_{h1}(nT) \ , \ R_{h2}(nT) \ , \ldots , \ R_{hk}(nT) \right]$$

où R_{h1} (n T) est le nombre d' intervalles entre deux passages à zéro à la sortie du filtre h , appartenant au groupe 1 pour le segment n ; k est le nombre total des groupes .

Une autre application de la méthode des passages à zéro est dans le cadre de la segmentation de la parole . Nous voyons sur la figure (3.24.) la variation du taux des passages à zéro sur un mot donne par ITO et DONALDSON / 19 /

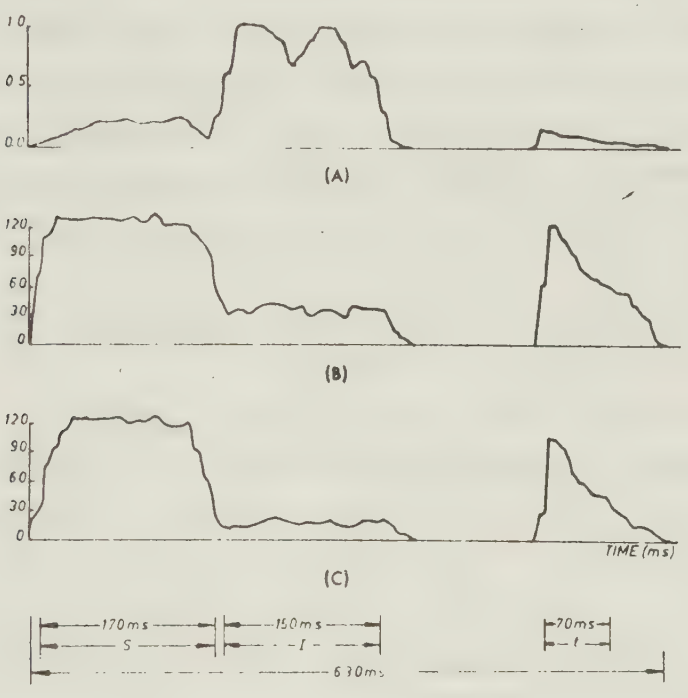


Figure 3.24. : A. Variation de l'amplitude sur une phrase
B. Variation du taux des passages à zéro de la dérivée du signal .
C. Variation du taux des passages à zéro du signal vocal (figure tirée de /19/)

Les limitations de cette méthode sont les suivantes :

a) la representativité par les taux des passages à zero peut souvent être mise en cause même pour les sons sonores
b) il y a beaucoup de difficultés pour représenter les sons sourds

Conclusion

Il a été prouvé que les techniques fréquentielles basées sur la deconvolution du signal vocal en utilisant le cepstre sont très intéressantes pour l'analyse et le synthèse de la parole . Le calcul de la transformée en z sur des contours passant plus près des pôles a d'ailleurs permis , en réduisant la bande passante , d'avoir des pics beaucoup plus décisifs pour les formants en assurant ainsi une détection plus facile . Ceci a donné la possibilité de lever certaines ambiguitées dans le cas des formants très rapprochés et a permis l'automatisation des procédés . Un modèle du spectre a été utilisé par ailleurs comme une autre solution pour faciliter la détection des formants . Les algorithmes établis peuvent donc être automatisés . Pourtant la décision sonore-sourd pour des raisons qui ont été déjà expliquées n'est pas toujours assurée et la mesure du pitch peut ne pas être toujours correcte .
En outre les méthodes sont par définition asynchrones et par conséquent les paramètres calculés sont moyennés.
Il faut ajouter qu'au point de vue " temps de calcul " , ces méthodes dépassent beaucoup le temps réel .(Plusieurs fois temps réel).
Nous voulons ajouter par la suite une remarque à laquelle on attache une certaine importance . Les quelques échantillons près de l'origine du cepstre (C_i) correspondant bien à la réponse impulsionnelle du canal vocal ; une fonction de transfert en z pourrait caractériser cette réponse impulsionnelle . Cette fonction de transfert est la suivante

$$\overset{*}{H}(z) = C_0 + C_1 z^{-1} + C_2 z^{-2} + \ldots + C_m z^{-m}$$

m étant de l'ordre de 30 .

Une fonction de transfert contenant donc uniquement des zéros représente le système de phonation . Mais en réalité les zéros sont rares et c'est plutôt des pôles qui apparaissent . Donc $\overset{*}{H}(z)$ approxime une fonction du type :

$$H(z) = \frac{1}{a_0 + a_1 z^{-1} + a_2 z^{-2} + \ldots + a_k z^{-k}}$$

Vu qu'il s'agit d'une approximation des pôles par des zéros, il faut s'attendre à ce que l'ordre du système $\overset{*}{H}(z)$ soit beaucoup trop élevé par rapport à la réalité. En effet, nous savons que (4.5) formants représentent complètement le processus de phonation, donc l'ordre effectif du système ne doit pas excéder 10 = (5 x 2).
Enfin, il faut ajouter que la fonction de transfert $\overset{*}{H}(z)$ ne correspond pas à une fonction "modèle" préétablie et recherchée mais à une fonction dérivée a posteriori d'un calcul tout à fait naturel qui ne correspond absolument pas à une optimisation. Cette réduction de redondance "naturelle" sans modèle et sans critère d'optimalité nous paraît très remarquable.

4. RECHERCHE DE NOUVELLES METHODES

4.1. Paramétrisation de la parole en utilisant un modèle de la membrane basilaire.

La plupart des méthodes utilisées jusqu'à présent, pour l'analyse du signal vocal , font appel en général à des modèles du système de phonation . Ainsi au chapitre précédent nous avons vu des méthodes fréquentielles et temporelles qui cherchaient à mesurer les fréquences de résonance du canal vocal . Nous verrons aux chapitres 5,6,7. des méthodes d'identification du processus de phonation en utilisant un modèle du conduit vocal .

MAISSIS /20/ au Laboratoire d'Automatique de l'E.N.S.T. a utilisé un modèle plutôt "acoustique ". Ses travaux sur l'analyse du signal vocal par un système régi par des équations aux dérivées partielles (modèle approximatif du comportement de l'oreille) lui ont permis d'engendrer des paramètres originaux dotés de propriétés intéressantes . Nous allons décrire par la suite ce système de paramétrisation , basé sur les propriétés de la membrane basilaire .
Comme on le voit sur la figure 4.1., la paramétrisation comprend les opérations suivantes :

1. Le signal temporel est d'abord segmenté en éléments minimaux . C'est une durée de 10 ms qui est adoptée.

2. Les passages à zéro du signal préfiltré dans les bandes :
(150 - 900) Hz , (900 - 2200) Hz , (2200 - 5000) Hz sont comptés .
Nous avons ainsi une estimation des trois premiers formants . Dans une version posterieure (/36/ , /59/) au lieu de trois premiers formants , sept fréquences dominantes et sept attenuations sont calculées.

3. Les paramètres proprement dits sont déterminés par la suite en utilisant une batterie de filtres passe-bande à large bande passante modelée sur les propriétés de la membrane basilaire . C'est ce point qui sera exposé par la suite .

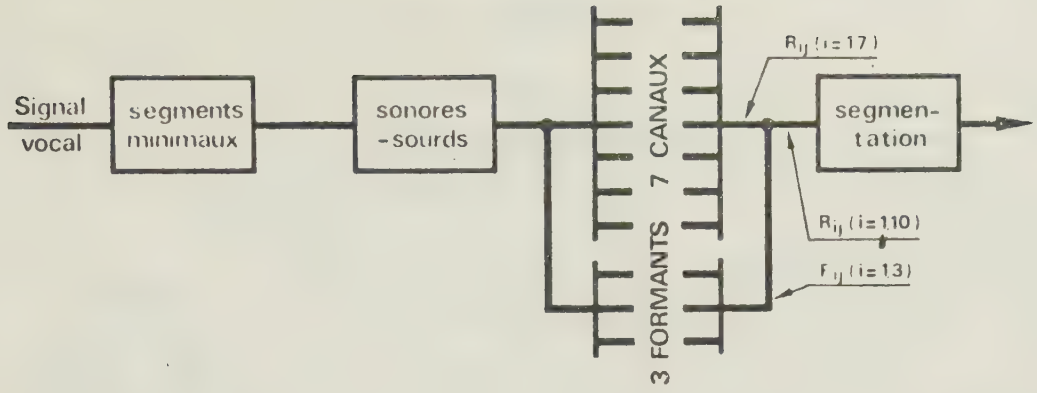


Figure 4.1 Système de paramètrisation basé sur les propriétés de la membrane basilaire.

Il est connu /3/ qu'un modèle approximatif de la réponse d'un point de la membrane basilaire (déplacements verticaux X(t,1) à la distance 1 de la fenêtre ovale) à l'excitation par un signal sonore peut être simulé par un filtre à large bande caracterisé par:

. Une fréquence de résonance dépendant du point considéré.
. Une bande passante à 3db proportionnelle à cette fréquence.
. Un gain maximal caractéristique du point étudié.

La méthode de parametrisation exposée utilise la reponse temporelle X(t,1) de 7 points de la membrane simulée et leurs dérivées spatiales définies comme suit.

a. Les 7 canaux admettent pour fréquences de résonance:
300 , 470 , 800 , 1300 , 2100 , 3600 , 4600 Hz de façon à recouvrir une partie importante du spectre de la parole.

b. La dérivée spatiale définie comme $\frac{\partial X}{\partial l}(t, l)$ nécessite la création d'une deuxième série de 7 canaux de fréquences 260 , 410 , 720 1200,1900 , 3350 , 4200 Hz correpondant à $\partial l \triangleq 0.5$ mm.

La figure 4.2. illustre le mode d'extraction des paramètres relatifs au canal k . Pour un segment minimal j , on détermine le maximum du signal à la sortie du filtre de type (a) , soit A (k ,j) et le minimum de la différence des sorties des filtres (a) et (b) correspondants , soit A' (k,j) . Le paramètre résultant est :

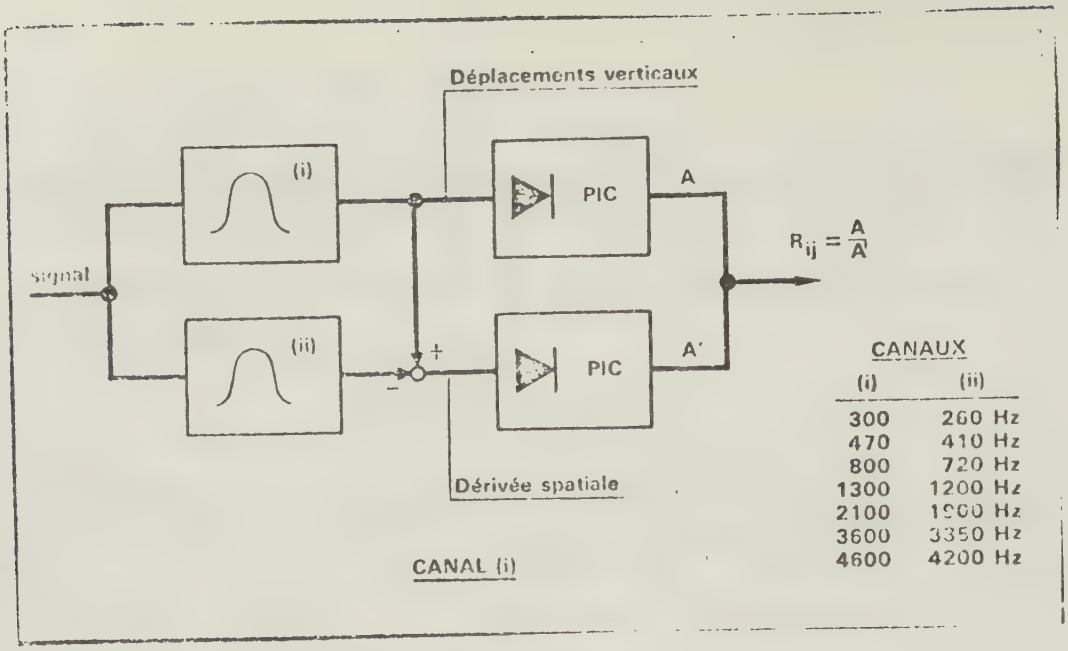


Figure 4.2. Extraction des paramètres basés sur un modèle de réponse de la membrane basilaire .

$$R(k, j) = \frac{A(k, j)}{A'(k, j)}$$

Pour N segments minimaux , le signal temporel est ainsi représenté par une matrice $R(k, j)$ de dimensions$(7, N)$

L'évolution du rapport $R(k, j)$ par rapport à j fait apparaître des régions de stabilité séparées par des régimes transitoires.

Par conséquent ces paramètres peuvent constituer un outil de segmentation. La figure 4.3. montre la variation du paramètre R_3 sur un mot .

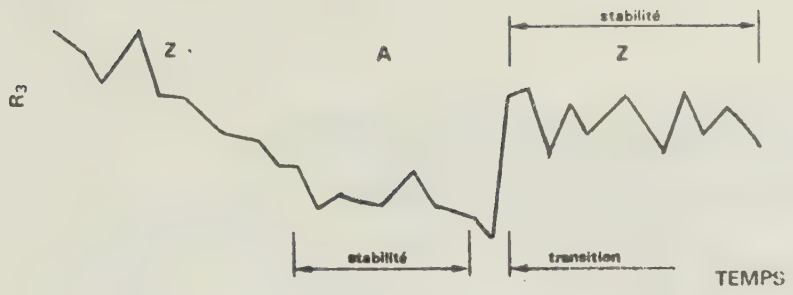


Figure 4.3. : Variation du paramètre R_3 sur le mot /ZAZOU/

En vue d'une utilisation de ces paramètres pour la reconnaissance des phonèmes ,deux sortes de normalisation sont effectuées : (/23/)

- Normalisation par rapport aux occurences
- Normalisation par rapport aux locuteurs.

On n'expliquera pas ici les détails de ces normalisations

4.2. Application à la reconnaissance des phonèmes.

Tout problème de reconnaissance des formes est constitué des deux étapes .

- Apprentissage
- Décision

Au niveau de l'apprentissage, le problème posé est d'effectuer un tri parmi les phonèmes inconnus caractérisés par les paramètres R (k ,j). Le but de ce tri est de grouper les phonèmes en classes symbolisées par le même concept abstrait. Avec la classification une réduction du nombre des paramètres est effectuée ; ceci pour réduire la mémorisation des formes d'apprentissage à celle d'un nombre restreint de paramètres . Ainsi l'approche cherchée est de représenter les paramètres dans un espace réduit d'une manière optimale au sens d'un critère donné . C'est l'analyse factorielle des correspondances qui est utilisée . L'analyse factorielle fournit dans un espace réduit de dimension fixée, la meilleure approximation inertielle du nuage d'apprentissage au sens d'une métrique distributionnelle du χ^2 sur l'ensemble des phonèmes pondérés .
Au niveau décision plusieurs méthodes ont été essayées . /36/ ,/59/ .

5. ANALYSE DE LA PAROLE EN UTILISANT UN MODELE IDENTIFICATION DU SYSTEME DE PHONATION

5.1. Définitions

Les méthodes fréquentielles d'analyse de la parole , ont connu beaucoup de succès et ont donné un moyen de représentation fidèle du signal vocal caractérisée par une grande diminution de la redondance. Les quatre dernières années , de nouvelles techniques ont été developpées pour analyser la parole directement dans le domaine temporel .
Dans le cadre de ces techniques on ne classera pas celles des passages à zéro , etc, bien qu'il s'agisse encore de techniques temporelles.
On a consacré un autre chapitre à ces techniques. Il s'agira ici des méthodes qui utilisent un modèle pour représenter la fonction de transfert en z du canal vocal .
Cette fonction de transfert est du type :

$$H(z) = \frac{b_o}{a_o + a_1 z^{-1} + \ldots\ldots + a_k z^{-k}} \qquad (5.1.)$$

Ce modèle est utilisé pour approximer les caractéristiques de transmission du canal vocal . L'ajustement du modèle est réalisé au sens d'un critère d'optimisation en utilisant les mesures du signal vocal avec un algorithme d'identification . Nous savons qu'identifier c'est tenter de donner une représentation mathématique à un processus réel , en utilisant des mesures de l'entrée et de la sortie du système associé . Il faut en plus un modèle qui puisse approximer suffisement bien le système . Dans le cas général l'entrée et la sortie sont connues à un bruit près (perturbation)
Dans notre cas , le système est le canal vocal (système de phonation).
L'entrée est l'excitation et la sortie le signal vocal . Le modèle utilisé sera un modèle plutôt abstrait d'une fonction de transfert en z dépourvue de zéros comme elle est écrite en (5.1.) . Ceci n'est pas en contradiction avec le caractère résonant du canal vocal et avec tout ce qui a été prouvé jusqu' alors en utilisant d'autres méthodes d'analyse .
La procédure d'identification qui sera suivie est donnée schématiquement sur la figure (5.1.)

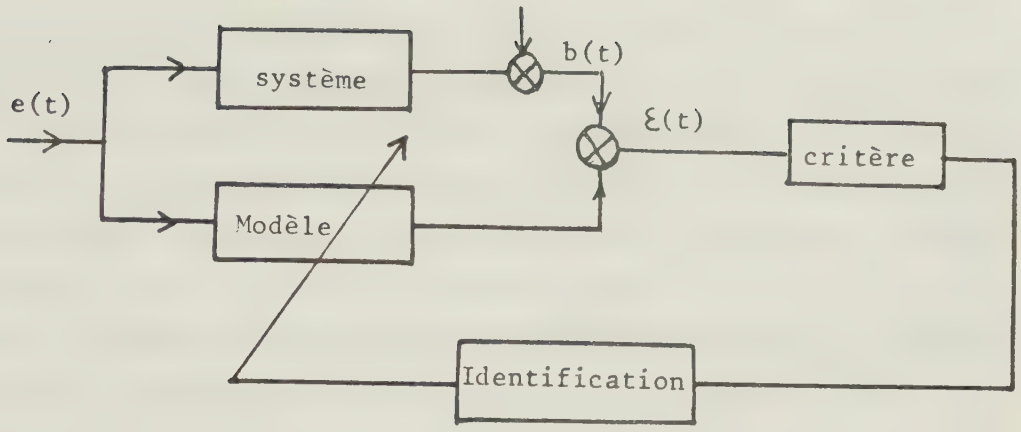


Figure 5.1. : Identification d'un système avec un modèle.

Pourtant l'entrée du système est inconnue et l'on pourrait difficilement concevoir une identification sans connaître l'entrée ; mais certaines suppositions peuvent être faites concernant cette entrée .
En particulier, comme nous l'avons vu au chapitre 1, l'excitation du canal vocal est soit du type impulsionnel , soit un bruit blanc . Dans ce cas notre problème peut avoir une solution comme on verra au chapitre suivant.

5.2. Modèlisation

Un certain nombre d'échantillons du signal vocal caractérisent comme le processus que l'on cherche à identifier.
Soit $\underline{s}^T = [s_o\ s_1 \ldots\ldots\ s_{N-1}]$ ce processus . Nous supposons en outre que $s_i = 0 \quad \forall i < 0$ et $\forall i \geqslant N$.

Le choix d'un modèle adéquat est toujours une étape importante dans la résolution d'un problème d'identification . Dans le cas du système de phonation , un grand nombre de modèles paramètriques plus ou moins réalistes au point de vue physiologique ont été proposés /1/,/3/ . On adoptera intentionnellement ici un modèle de forme très générale (fonction de transfert en z) car c'est plus la caracterisation abstraite du signal qui est recherchée ici que celle des organes de phonation . Un tel modèle couramment utilisé est celui proposé par FANT /1/ qui considère le signal vocal comme la réponse d'un système acoustique (conduit vocal , dérivation nasale , radiation buccale ...) à une excitation de type fixé . Dans le cas des sons sonores , l'entrée est ainsi constituée par une série d'impulsions quasi périodiques émanant des cordes vocales, et dans le cas des sons sourds , par un bruit blanc dû aux turbulences de l'air dans les constrictions du canal vocal . Les deux sources d'excitation peuvent de plus coexister. L'identification consiste alors à ajuster les paramètres du modèle pour le rendre aussi voisin que possible du système réel au sens d'un critère donné : similitude des sorties /10/, de leurs spectres /5/ ....

Supposant le système de phonation linéaire , il pourra être représenté pour les nécessités du traitement numérique par une équation recurrente d'ordre k , liant les valeurs échantillonnées de l'entrée e_n (excitation) et de la sortie s_n (signal vocal) :

$$a_o s_n + a_1 s_{n-1} + \ldots\ldots + a_k s_{n-k} = b_o e_n + b_1 e_{n-1} + \ldots + b_k e_{n-k} \qquad (5.2.)$$

soit, appliquant la transformée en z (conditions initiales nulles):

$$S(z) = H(z).E(z) \quad \text{avec} \quad H(z) = \frac{b_o + b_1 z^{-1} + \ldots\ldots + b_k z^{-k}}{a_o + a_1 z^{-1} + \ldots + a_k z^{-k}} \qquad (5.3.)$$

Ce modèle peut être particularisé en assimilant le canal vocal à une serie de résonnateurs en cascade. Pour les sons non-nasalisés , la fonction de transfert H(z) ne comporte alors que des pôles . La mise en oeuvre de la dérivation nasale introduit des zéros dans H(z) mais ceux-ci (sous réserve de stabilité) peuvent être approximés par un nombre convenable de pôles . Il en résulte qu'un modèle suffisant est caractérisé par $b_i = o$ pour $i \neq o$ dans (5.2) et (5.3.)

La relation entrée-sortie peut être explicitée , en l'absence de conditions initiales , sous forme matricielle :

$$b_o \begin{bmatrix} e_1 \\ e_2 \\ \cdot \\ \cdot \\ e_k \\ \cdot \\ \cdot \\ e_N \end{bmatrix} = a_o \begin{bmatrix} s_1 \\ s_2 \\ \cdot \\ \cdot \\ s_k \\ \cdot \\ \cdot \\ s_N \end{bmatrix} + \begin{bmatrix} s_o & o & \cdots & o \\ s_1 & s_o & \ddots & \vdots \\ \cdot & \ddots & \ddots & o \\ s_{k-1} & \cdots & \cdots & s_o \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ s_{N-1} & \cdots & \cdots & s_{N-k} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ a_k \end{bmatrix} \qquad (5.4.)$$

$$b_o \underline{e} = a_o \underline{s} + S \underline{a} \qquad (5.5)$$

où S est une matrice de Hankel à k colonnes .

Pour tenir compte d'éventuelles conditions initiales , il suffit de n'écrire la relation (5.4.) qu'à partir de l'indice k.La matrice S est alors complète et s_i,(i=0,...,k-1) jouent le rôle de conditions initiales Le choix d'un tel modèle résulte d'un compromis entre l'exactitude et la complexité de la représentation.La complexité du processus de phonation

introduit certaines difficultés dans son identification :

i- L'entrée du système n'est connue que très approximativement (le modèle bruit-impulsion est un peu simpliste) Un certain nombre de sources intermédiaires coexistent à l'intérieur même du canal vocal constrictions en particulier).

ii- Le système est variable dans le temps et en conséquence la période d'identification nécessairement courte . Le canal vocal est un conduit acoustique à paramètres répartis dont on cherche à rendre compte globalement par un modèle linéaire entrée-sortie . L'ordre de ce modèle dépendant de la propagation des ondes dans ce conduit , est donc essentiellement variable .

iii-Le signal acoustique capté est toujours entaché de bruit , ce qui limite la précision de l'identification et , pour les ordres élevés , augmente la variabilité du modèle .

Si l'on suppose par ailleurs , que le système est stable et que sur l'intervalle (0,N-1) , il n'y a aucune intervention de l'entrée $e_i = o$, $i = o , N - 1$, on peut admettre que $s_i = o$ pour $i > N - 1$ et la matrice S se terminera par une partie triangulaire inférieure limitée par S_{N-1} . Ceci revient à imposer des conditions finales sur le processus .

On appelle S^+ la matrice S augmentée pour tenir compte de ces conditions finales .

$$
S^{+} = \begin{bmatrix}
s_0 & 0 & \cdots & 0 & 0 \\
s_1 & s_0 & \cdots & \cdots & \cdot \\
\cdot & \cdot & \cdots & \cdots & \cdot \\
\cdot & \cdot & \cdots & \cdots & 0 \\
s_{k-1} & \cdot & \cdots & \cdots & s_0 \\
\cdot & \cdot & \cdots & \cdots & \cdot \\
\cdot & \cdot & \cdots & \cdots & \cdot \\
\cdot & \cdot & \cdots & \cdots & \cdot \\
s_{N-1} & \cdot & \cdots & \cdots & s_{N-k} \\
0 & s_{N-1} & \cdot & \cdots & s_{N-k+1} \\
\cdot & \cdot & \cdot & \cdots & \cdot \\
\cdot & \cdot & \cdot & \cdots & \cdot \\
0 & \cdot & \cdot & 0 & s_{N-1}
\end{bmatrix}
$$

Nous tâcherons par la suite de présenter un modèle plus général , celui d'une fonction de transfert contenant des zéros . Ceci pour montrer les difficultés qu'une telle identification présenterait . Un tel modèle est décrit par 1'équation (5.3.)
L'équation récurrente correspondante s'écrit :

$$a_o s_n = a_1 s_{n-1} + a_2 s_{n-2} + \cdots + a_k s_{n-k} - b_o e_n - b_1 e_{n-1} - \cdots - b_r e_{n-r}$$

et sous forme matricielle pour n = 0 , ... , N :

$$-a_0 \begin{bmatrix} s_0 \\ s_1 \\ s_2 \\ . \\ . \\ . \\ s_N \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & . & . & . & 0 & e_0 & 0 & 0 & . & . & . & 0 \\ s_0 & 0 & 0 & . & . & . & 0 & e_1 & e_0 & 0 & . & . & . & 0 \\ s_1 & s_0 & 0 & . & . & . & 0 & e_2 & e_1 & e_0 & . & . & . & 0 \\ . & . & . & . & . & . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . & . & . & . & . & . \\ s_{N-1} & s_{N-2} & . & . & . & . & s_{N-k} & e_N & e_{N-1} & . & . & . & . & e_{N-r} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ . \\ . \\ . \\ -b_0 \\ . \\ . \\ . \\ -b_r \end{bmatrix}$$

A partir de cette équation on peut s'apercevoir de l'impossibilité d'identification si l'on ne dispose pas des mesures de l'entrée La seule solution pour l'application de ce modèle généralisé consiste à choisir des coefficients b_i (i= 1 ,..., r) constants (connus ou précalculés). Dans ce cas , ce sont seulement les coefficients a_i qui sont ajustés par l'algorithme d'identification.

6. METHODES GLOBALES D'IDENTIFICATION

Nous utiliserons le terme "méthodes globales" pour les méthodes qui cherchent à identifier un modèle du système de phonation en tenant comte du processus défini sur un horizon déterminé. Le signal est ainsi consideré d'une manière globale par sa fonction d'autocorrelation. C'est la résolution d'un système d'équations linéaires (équations d'autocorrelation) qui conduit à la solution du problème.

6.1 Prédiction linéaire par ATAL- HANAUER

ATAL et SCHROEDER /25/ et ensuite ATAL et HANAUER /10/ ont présenté un algorithme d'analyse de la parole basé sur un modèle comme celui du chapitre précedent. Le schéma fonctionnel de la figure 6.1 est utilisé pour répresenter le modèle recurrent.

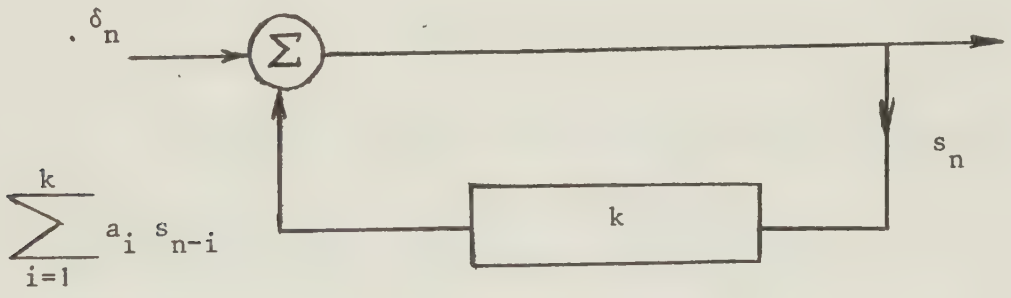


Figure 6.1 : Schéma fonctionnel

Pour les sons sonores tous les échantillons du processus ont été considérés prédictibles sauf un seul échantillon au début de chaque période (entrée strictement impulsionnelle). Donc la relation suivante est utilisée pour les sons sonores:

$$\hat{s}_n = \sum_{i=1}^{k} a_i \, s_{n-i}$$

Pour les sons sourds l'entrée est approximée par un bruit pseudo-blanc δ_n :

$$\hat{s}_n = \sum_{i=1}^{k} a_i \, s_{n-i} + \delta_n$$

La minimalisation se porte sur l'erreur quadratique moyenne entre la sortie réelle et la sortie estimée (critère des moindres carrés)

La quantité à minimaliser est donc :

$$\varepsilon^2 = \sum_{n=1}^{N} \left(s_n - \hat{s}_n \right)^2$$

$$\varepsilon^2 = \sum_{n=1}^{N} \left(s_n - \sum_{i=1}^{k} a_i \, s_{n-i} \right)^2$$

$$\varepsilon^2 = \left(\underline{s} - S\underline{a} \right)^T \left(\underline{s} - S\underline{a} \right) \qquad (6.3.)$$

L'estimation optimale $\hat{\underline{a}}$ est alors donnée en prenant :

$$\frac{\partial \varepsilon^2}{\partial a} = 2 \left(S^T \underline{s} + S^T a \right) = 0$$

D'où la solution définie par

$$\left[S^T S \right] \hat{\underline{a}} = . \; S^T \underline{s} \qquad (6.4.)$$

d'où $\hat{\underline{a}} = \left[S^T S \right]^{-1} S^T \underline{s}$

On pose $S^* = \left[S^T S \right]^{-1} S^T$

Ainsi $\hat{\underline{a}} = S^* \underline{s}$ S^* est la pseudo -inverse de S

L'erreur minimale correspondante sera :

$$\bar{\varepsilon}^2_{min} = \underline{s}^T \left(I - S \left(S^T S \right)^{-1} S^T \right) \underline{s} \qquad (6.5.)$$

Dans le cas des sons sourds , on peut supposer que l'entrée est un bruit pseudo blanc de moyenne nulle . La solution de (6.4.) minimalise alors l'erreur de prédiction considérée comme aléatoire ($E(\varepsilon^2)$ minimale)

Méthode de résolution.

L'équation 6.4. peut s'écrire :

$$\boxed{V\underline{a} = \underline{v}} \qquad (6.6.)$$

où
$$V = S^T S$$
$$\underline{v} = S^T \underline{s} \qquad (6.7.)$$

$$V = \begin{bmatrix} v_{00} & v_{10} & v_{20} & \cdots & v_{k-1,0} \\ v_{10} & v_{11} & v_{12} & \cdots & \\ v_{20} & v_{21} & v_{22} & \cdots & \\ \vdots & \vdots & \vdots & & \vdots \\ v_{k-1,0} & & & \cdots & v_{k-1,k-1} \end{bmatrix} \qquad (6.8.)$$

L'élément v_{ij} de la matrice V peut être calculé à partir de la relation:

$$v_{ij} = \sum_{n=0}^{N-1-/i/} s_n \, s_{n+|i-j|} \qquad (6.9.)$$

Le vecteur $\underline{v}$ est constitué à partir des éléments suivants :

$$\bar{v} = (v_{10} \quad v_{20} \quad \ldots\ldots \quad v_{k0})$$

En ce qui concerne la matrice V on remarque qu'elle est symètrique . Elle est de plus définie positive . En effet la dérivée seconde de (6.3.) donne

$$\frac{\partial^2 \varepsilon^2}{\partial a^2} = 2 S^T S = 2V > 0 \qquad \text{(pour avoir un minimum)}$$

La matrice V étant symétrique définie positive,
La résolution de l'équation 6.4. peut s'effectuer en utilisant la méthode de factorisation de V en deux matrices triangulaires (supérieure et inférieure) (Cholensky)

$$V = S^T S = R R^T \qquad \text{(R : matrice triangulaire)} \qquad (6.10.)$$

$$R = \begin{bmatrix} r_{11} & 0 & \cdots & 0 \\ r_{12} & r_{22} & & \vdots \\ \vdots & & & 0 \\ r_{1k} & r_{2k} & \cdots & r_{kk} \end{bmatrix} \qquad (6.11.)$$

Si Δ_i sont les mineurs premiers de $\left[S^T S\right]$

c'est à dire $\Delta_0 = 1$, $\Delta_1 = v_{00}$, $\Delta_2 = \begin{vmatrix} v_{00} & v_{10} \\ v_{10} & v_{11} \end{vmatrix}$, etc

Il est interessant de noter que :

$$r_{11}^2 = \frac{\Delta_1}{\Delta_0} \qquad r_{22}^2 = \frac{\Delta_2}{\Delta_1} \quad \cdots\cdots\cdots \qquad r_{kk}^2 = \frac{\Delta_k}{\Delta_{k-1}}$$

et que cette méthode de factorisation de matrice correspond à la construction d' une base canonique de la forme quadratique $V(\underline{a} ; \underline{a}) = \underline{a}^T V \underline{a}$ (6.13.)

on aura donc $V(\underline{a};\underline{a}) = \frac{\Delta_0}{\Delta_1} f_1^2 + \frac{\Delta_1}{\Delta_2} f_2^2 + \ldots + \frac{\Delta_{k-1}}{\Delta_k} f_k^2$

où $(f_1, f_2, \ldots, f_k)$ sont les coordonnées de $\underline{a}$ par rapport à la nouvelle base

Les valeurs r_{ii} ont aussi une interprétation géométrique intéressante sur le plan des f_i (voir figure 6.2.)

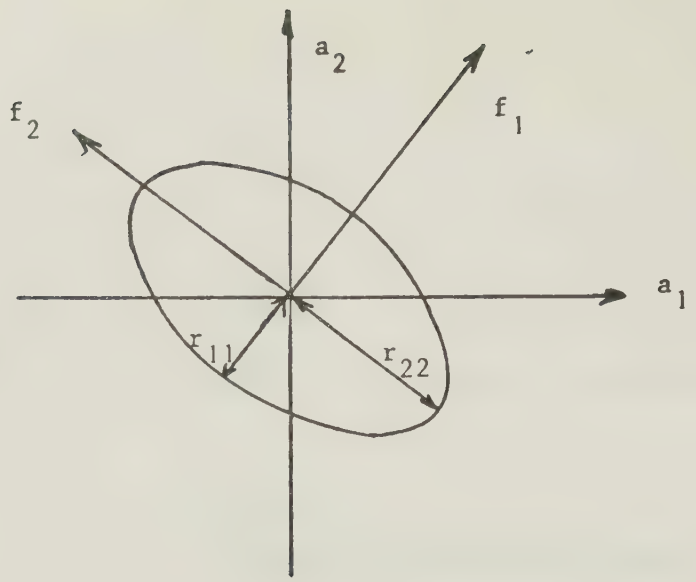


Figure 6.2. : Interprétation géométrique des valeurs r_{ii}.

Elles représentent les axes de l'ellipsoïde : $A(\underline{a} ; \underline{a}) = 0$ (voir / 26 /.)

La résolution du système après factorisation est simple

Il suffit d'écrire

$$R\underline{x} = v \qquad (6.15.)$$

$$R^T\underline{a} = \underline{x}$$

La résolution de ces deux systèmes est facile

En ce qui concerne le temps de calcul :

Il faut environ (N.k) multiplications pour former la matrice $\left[S^T S\right]$

Il faut remarquer qu'il suffit de calculer la 1ère ligne (ou 1ère colonne) de cette matrice et que les autres termes s'en déduisent à un terme correctif près .

Finalement le calcul de $\hat{\underline{a}}$ nécessite :

$k\left(N + \frac{k^2}{6} \right)$ multiplications + k racines carrées . Le temps disponible pour effectuer ces opérations est de l'ordre de la durée T_o de la période fondamentale . En ce qui concerne la mémoire nécessaire , il faut noter que seuls la matrice $\left[S^T S\right]$ et le vecteur $\underline{v}$ doivent être disposés en mémoire centrale . Donc le nombre de mémoires à réserver est indépendant de N ; en outre vu les symétries de $\left[S^T S\right]$ le nombre de mots à réserver est

$k (k + 5) / 2$

Une dernière remarque est relative au conditionnement de la matrice $\left[S^T S\right]$

Il faut remarquer que $\xi(S^T S) = \left[\xi(S) \right]^2$ ξ: nombre de condition

par conséquent la méthode sera peu performante si la matrice S est mal conditionnée.

Concernant les temps de calcul ATAL a donné les indications suivante

Sur GE635 (Cycle mémoire $1\,\mu s$)
Sans préciser le langage de programmation
pour k = 10 et T = 100 μs

Opération	temps de calcul
Extraction des coefficients de prédiction à partir de 100 échantillons	75 m sec / segment
Calcul de l'enveloppe du spectre (500 échantillons du spectre)	250 m sec / segment
Fréquences des formants et amortissements.	60 ms / segment
k échantillons de la fonction d'autocorrelation à partir des coefficients de prédiction	10 ms / segment

Parole Synthetique à partir des coefficients de prédiction .	8 fois temps réel
Pitch	10 fois temps réel

6.2. Méthode de maximum de vraisemblance

SAITO et ITAKURA /27/,/28/ ont abouti ,par des considérations différentes , à une équation de covariance à résoudre , du même type que celle d'ATAL-HANAUER . Un postulat supplémentaire a été utilisé , celui du processus Gaussien stationnaire pour le signal vocal . Les paramètres $\underline{r} = (\sigma^2 , a_1 , a_2 , \ldots , a_k)$ d'une fonction de transfert du type:

$$H(z) = \frac{\sigma^2}{2\pi} \cdot \frac{1}{a_0 + a_1 z^{-1} + \ldots + a_k z^{-k}}$$

sont recherchés. Soit $\underline{s} = (s_1 , s_2 , \ldots , s_N)$ le processus considéré.On choisira comme estimation de $\underline{r}$, celle qui maximalise la fonction de vraisemblance. Cette fonction est en général définie par la densité conditionnelle de probabilité de $\underline{s}$ étant donné $\underline{r}$; soit:

$$L(\underline{s}) = \text{Log } p(\underline{s} / \underline{r}) \qquad (6.16)$$

où $p (\underline{s} / \underline{r}) = \dfrac{C}{(2\pi)^{N/2} \left| V \right|^{1/2}} \exp (- \dfrac{1}{2} \underline{s}^T V \underline{s})$

V est une matrice (N × N) avec des éléments $v_{ij} = E (s_i s_j)$ et $\left| V \right| = \det (V)$

La maximalisation de p ($\underline{s}$ / $\underline{r}$) conduit à un système d'équations du type :

$$\sum_{l=0}^{k} a_l \, v_{l-i} = 0 \qquad i = 1 , 2 , \ldots\ldots k \qquad (6.16.)$$

et

$$\sum_{l=0}^{k} a_l \, v_l = \sigma^2 \qquad (6.17.)$$

soit sous une autre forme :

$$V \, \underline{a} = \underline{v} \qquad (6.18.)$$

et $$\underline{a}^T \underline{v} = \sigma^2$$

où

$$V = \begin{bmatrix} v_o & v_1 & v_2 & \ldots\ldots \\ v_1 & v_o & v_1 & \ldots\ldots \\ \ldots\ldots & & & \\ \ldots\ldots & & & \end{bmatrix} \qquad (6.19.)$$

et $$v_k = \sum_{n=0}^{N-1-|k|} s_n \, s_{n+|k|} \qquad (6.20.)$$

La démonstration de(6.18) se trouve dans /27/et avec un formalisme différent dans/29 /

Il faut remarquer que la matirce V utilisée ici est légèrement différente de la matrice V utilisée par ATAL-HANAUER . Il s'agit d'une petite différence (quelques termes additifs) dans la définition des v_i . La relation entre les v_{ik} de la première méthode et les v_i de la seconde est la suivante :

$$v_i = v_{ik} + \frac{\sum_{N-1-|i|}}{N-|i|-k} s_n s_{n+|i|} \qquad (6.21.)$$

Mais il faut ajouter que dans le cas d'un traitement asynchrone où une fenêtre de Hamming est utilisée pour pondérer le processus la valeur de la somme de cette relation est très petite, on aura, alors, $v_i \cong v_{ik}$ et par conséquent les matrices V sont pratiquement les mêmes . Donc avec des considérations tout à fait différentes on a abouti à la même équation d'autocorrélation à résoudre que celle utilisée par la méthode précédente , avec une légère différence pour la définition des termes d'autocorrelation v_i . Il serait interessant par la suite de voir si on pourrait abouti à la même relation avec un formalisme de moindres carrés . La réponse est positive . Il suffit de considérer au lieu de la matrice S la matrice S^+ qui tient compte des conditions finales du processus comme ceci a été déjà expliqué au CH 5.2.
A l'aide de cette matrice on aboutit, en utilisant un critère des moindres carrés, à la même équation que ITAKURA - SAITO avec une légère différence dans la définition des v_i.

Résolution de l'équation d'autocorrélation

Nous exposons ici une méthode de résolution basée sur la définition d'un produit scalaire telle qu'elle a été développée par MARKEL /29/. Nous gardons excactement la même formulation que MARKEL. Cette méthode a été utilisée par ITAKURA - SAITO /28/. Nous posons :

$$A(z) = 1 + a_1 z^{-1} + \ldots + a_k z^{-k} = \sum_{i=0}^{k} a_i z^{-i} \qquad (6.22.)$$

avec : $a_o = 1$

On définit les polynômes :

$$F(z) = \sum_{k=-\infty}^{+\infty} f_k z^{-k} \qquad G(z) = \sum_{k=-\infty}^{+\infty} g_k z^{-k} \tag{6.23.}$$

$$F^+(z) = \sum_{k=-\infty}^{+\infty} f_k^* z^k \qquad V(z) = \sum_{k=-\infty}^{+\infty} u_k z^{-k}$$

f_k^* est le conjugué de f_k .

Pour f_k réel $\quad F^+(z) = F(1/z)$ (6.24.)

Si $V(z)$ est la transformée en z de la séquence de correlation v_k en appliquant l'intégrale de la transformée inverse , on aura :

$$v_k = v_{-k} = \frac{1}{2\pi j} \int_\Gamma V(z)\, z^{-k-1}\, dz \qquad (j = \sqrt{-1}) \tag{6.25.}$$

où Γ est un contour fermé qui contient l'origine du plan des z . Si Γ est le cercle unitaire :

$$v_k = \frac{1}{2\pi} \int_{-\pi}^{+\pi} V(e^{j\theta})\, e^{-jk\theta}\, d\theta \tag{6.26.}$$

Nous définissons un produit scalaire par :

$$\langle F(z) , G(z) \rangle = \frac{1}{2\pi j} \int_\Gamma V(z)\, F^+(z)\, G(z)\, z^{-1}\, dz$$

$$= \frac{1}{2\pi} \int_{-\pi}^{+\pi} V(e^{j\theta})\, F(e^{j\theta})\, G(e^{j\theta})\, d\theta \tag{6.27.}$$

Ceci étant on aura :

$$v_{i-j} = \langle z^{-i} , z^{-j} \rangle \qquad (6.28.)$$

Avec cette formulation on peut écrire les équations d'autocorrelation sous une autre forme :

$$\sum_{i=0}^{k} a_i \langle z^{-i} , z^{-j} \rangle = 0 \qquad j = 1,2....k \qquad (6.29.)$$

Soit : $$\langle \sum_{i=0}^{k} a_i z^{-i} , z^{-j} \rangle = 0 \qquad j = 1,2 ....k$$

or : $$\langle A(z) , z^{-j} \rangle = 0 \qquad j = 1,2 .....k \qquad (6.30.)$$

Cette dernière relation conduit à un résultat très important :
Les équations d'autocorrelation peuvent être résolues en forçant A (z) à être orthogonal à z^{-j} pour j = 1,2 .... k.

Pour démontrer que la définition 6.27 est la définition d'un vrai produit scalaire il suffit de démontrer les trois relations ci dessous :

(1) $$\langle F(z) , G(z) \rangle = \langle G(z) , F(z) \rangle^{*} \qquad (6.31.)$$

(2) $$\langle cF(z), aG(z) + b U(z) \rangle \quad a c^{*} \langle F(z) , G(z) \rangle + bc^{*} \langle F(z) , U(z) \rangle$$

(3) $$\langle F(z) , F(z) \rangle \quad \text{positif pour } F(z) \neq 0$$

En outre les relations suivantes peuvent être démontrées facilement .

$$\langle z^{-n} F(z) , z^{-n} G(z) \rangle = \langle F(z) , G(z) \rangle$$

$$\langle F(z) , G(z) \rangle = \langle 1, F^{+}(z) G(z) \rangle \qquad (6.32.)$$

$$\langle F(z) , G(z) \rangle = \sum_{m=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} f_m^* g_l v_{m-l}$$

Une procédure récurrente sera utilisée par la suite pour la résolution des équations d'autocorrélation basée sur la constatation (6 . 30.)
Soit $A_m(z)$ le filtre d'ordre m et $A_{m+1}(z)$ le filtre d'ordre (m+1) la méthode de résolution qu'on expose par la suite montrera comment peut se faire le passage de $A_m(z)$ à $A_{m+1}(z)$ en tenant compte des relations qu'on a évoquées précedemment . Le résultat final sera $A(z) = A_k(z)$ ($m = k$)

Pour le filtre d'ordre m nous écrirons :

$$A_m(z) = \sum_{i=0}^{m} a_{mi} z^{-i} \qquad \text{où } a_{mo} = 1 \quad \forall m \qquad (6.33.)$$

Nous allons alors montrer qu'on peut calculer un polynôme $B_m(z)$ d'ordre (m+1), fonction du polynôme $A_m(z)$ d'ordre m, qui est aussi orthogonal aux puissances de z^{-1}.

Nous savons que :

$$\langle A_m(z) , z^{-j} \rangle = \langle z^j , A_m(1/z) \rangle = \langle z^{-m-1+j} , z^{-m-1} A_m(\frac{1}{z}) \rangle = 0$$

pour $j = 1,2 ,.....m$

Nous définissons l'indice $i = -m -1 + j$

et le polynôme :

$$B_m(z) = z^{-(m+1)} A_m(1/z) \tag{6.34.}$$

On aura : $\langle z^{-i}, B_m(z) \rangle = 0 \qquad i = 1, 2, \ldots\ldots m$

où $B_m(z) = z^{-(m+1)} \sum_{i=0}^{m} a_{mi} z^i = \sum_{i=1}^{m+1} b_{mi} z^{-i}$ (6.35.)

$$b_{mi} = a_{m, m+1-i} \quad , \quad b_{m,m+1} = a_{mo} = 1 \tag{6.36.}$$

Jusqu'à présent deux polynômes $A_m(z)$ d'ordre m et $B_m(z)$ d'ordre (m+1) sont définis chacun orthogonal aux puissances de z^{-j} , $j = 1, \ldots. m$
Un polynôme $A_{m+1}(z)$ d'ordre (m+1) peut être défini tel que :

1. Le coefficient de z^o soit égal à 1
2. $A_{m+1}(z)$ soit orthogonal aux puissances de z^{-j} , $j = 1, \ldots., m, m+1$

Dans ce cas la procédure réccurente sera déterminée . Toute combinaison linéaire de $A_m(z)$ et $B_m(z)$ sera un polynôme d'ordre (m+1) puisque $B_m(z)$ est d'ordre (m+1) . Toute combinaison linéaire de $A_m(z)$ et $B_m(z)$ sera orthogonale aux puissances de z^{-j} de z^{-1} jusqu'à z^{-m} .
Vu que le coefficient de z^o est l'unité pour $A_m(z)$ et zéro pour $B_m(z)$, la combinaison linéaire :

$A_{m+1}(z) = A_m(z) + k_m B_m(z)$ où k_m = cte satisfait la première condition .
Pour satisfaire aussi la seconde il suffit que $A_{m+1}(z)$ soit orthogonal à $z^{-(m+1)}$

Si on définit :

$$\alpha_m = \langle z^{-(m+1)} , B_m(z) \rangle \qquad (6.38.)$$

$$\beta_m = \langle z^{-(m+1)} , A_m(z) \rangle \qquad (6.39.)$$

le produit $\langle z^{-(m+1)} , A_{m+1}(z) \rangle = 0$ donne

$$k_m = - \beta_m / \alpha_m \qquad (6.40.)$$

La procèdure récurrente est par conséquent déterminée. Le point de départ sera : $A_o = 1$

et les conditions initiales : $\alpha_o = \langle z^{-1} , B_o(z) \rangle = \langle z^{-1} , z^{-1} \rangle = v_o$

$$\beta_o = \langle z^{-1} , A_o(z) \rangle = \langle z^{-1} , z^{o} \rangle = v_1 \qquad (6.41.)$$

et $a_{oo} = 1$

A l'étape m, connaissant k_m, on détermine les coefficients de $A_{m+1}(z)$ comme ci dessous :

$$a_{m+1,\,l} = \begin{cases} a_{ml} = 1 & l = 0 \\ a_{ml} + k_m a_{m,\,m+1-l} & l = 1,2,\ldots,m \\ k_m & l = m+1 \end{cases} \qquad (6.42.)$$

Pour le calcul de k_m il faut calculer les α_m et β_m .
On peut avoir différentes expressions de α_m , β_m à partir des paramètres a_{mi} $i = 0,1....m.$
Par définition :

$$\alpha_m = \langle z^{-(m+1)} , B_m(z) \rangle = \langle z^{-(m+1)} , z^{-(m+1)} A_m(1/z) \rangle =$$

$$= \langle A_m(z) , 1 \rangle$$

Donc $$\alpha_m = \sum_{i=0}^{m} a_{mi} v_i \qquad (6.43.)$$

et pour β_m : $$\beta_m = \sum_{i=0}^{m} a_{mi} v_{m+1-i} \qquad (6.44.)$$

On peut avoir k_m à partir de α_m et β_m. Ces trois grandeurs sont des paramètres intermédiaires de calcul. A part, les formules précédentes, d'autres formules théoriquement équivalentes, peuvent être déduites. Ces formules peuvent donner des résultats numériquement différents si un ordinateur avec des mots de longueur différente est utilisée.

MARKEL a remarqué que les α_i et les k_i étaient des paramètres plus intéressants (dans le cadre de la transmission des données) que les a_i✱. Ceci est un grand avantage de la méthode et il peut aboutir à des réductions du taux de transmission importantes. La reconstruction des a_i à partir des k_i ou des α_i est simple.

Nous voulons remarquer ici un point commun entre cette méthode et la méthode utilisée par ATAL et expliquée précédemment. Avec la méthode du § 6.1, on avait comme résultat intermédiaire des calculs, les termes diagonaux de la matrice de covariance (après factorisation), qui étaient égaux à :

$$r_{11}^2 = \frac{\Delta_1}{\Delta_0} \quad , \quad r_{22}^2 = \frac{\Delta_2}{\Delta_1} \quad \ldots\ldots\ldots\ldots r_{kk}^2 = \frac{\Delta_k}{\Delta_{k-1}}$$

✱En effet, les a_i sont très sensibles à la quantification. Ainsi, on ne peut pas arriver à des réductions importantes du taux d'information à transmettre, en utilisant ces paramètres. Les α_i et les k_i sont moins sensibles à la quantification.

où les Δ_i sont les mineurs premiers de $S^T S$

Ici comme étape intermédiaire, on a eu les α_i , β_i , k_i .
Concernant les α_i , il faut noter que

$$\alpha_i = \frac{\Delta_{i-1}}{\Delta_i} \text{ où les } \Delta_i \text{ sont les mêmes mineurs}$$

donc : $$r_{ii}^2 = (1/\alpha_i) \qquad (6.45.)$$

On démontre ainsi que les deux algorithmes passent à un certain moment par le calcul de la même quantité. On pourrait donc se servir de cette quantité pour comparer leur qualité. On peut aussi chercher à quel stade de calcul la qualité se dégrade, etc ...

Une autre possibilité offerte par cet algorithme comme ceci a été remarqué par ITAKURA-SAITO et MARKEL est le contrôle de la stabilité. Ce contrôle peut être effectué à n'importe quel stade de la récurrence avec la valeur du coefficient k_m. Il est facile à démontrer que, si le calcul est exact, donc il n'y a pas d'instabilité, k_m est toujours tel que $|k_m| < 1$.

Par contre, si au moment du calcul $|k_m| > 1$ au stade m, ceci veut dire que le filtre d'ordre m est instable et que le filtre d'ordre (m + 1) sera instable. Par conséquent, le coefficient k_m est un bon moyen de constatation de l'instabilité en temps utile. Malheureusement, on reste à la constatation sans pouvoir remédier à l'instabilité.

On notera que malgré le caractère iteratif de l'algorithme de résolution (nombre de pas égal à k), la méthode n'est pas récurrente par nature. Le calcul nécessite, en effet, l'acquisition des N valeurs du signal s_n à la fois (où N couvre plusieurs périodes fondamentales). Il s'agit donc, comme pour ATAL et HANAUER, d'une méthode globale par opposition à celle proposée au Chapitre 7.

Pour conclure, la méthode de résolution basée sur l'algorithme de ROBINSON, est très prometteuse. Les avantages sont indiqués dans ce qui suit :

1. Temps de calcul très réduit donc possibilité de temps réel avec un hardware peu coûteux.
2. Contrôle de la stabilité (immédiat et dans le temps).
3. Transmission possible avec des paramètres intermédiaires moins sensibles donc possibilité de diminution du taux de transmission.

6.3 Filtrage inverse

Un filtre de fonction de transfert : $A(z) = \sum_{i=0}^{k} a_i \, z^{-i}$

est considéré comme le filtre inverse.

Etant donnée une séquence d'éntrée s_n , ($0 \leqslant n < N$) nous voulons déterminer les coefficients a_i de façon à minimiser l'énergie totale à la sortie du filtre .Cette énergie est calculée à partir d'une fonction d'erreur y_n telle que :

$$\varepsilon^2 = \sum_{n=0}^{L} y_n^2$$

où $$y_n = s_n + \sum_{i=1}^{k} a_i \, s_{n-i} \quad , (L = N + k - 1)$$

L'énergie totale peut devenir minimale pour des valeurs de a_i telles que les dérivées partielles de ε^2 par rapport à a_i,($i = 1, \ldots, k$) soient égales à zéro.

On aura ainsi :

$$\sum_{i=1}^{k} a_i \sum_{n=0}^{L} s_{n-i} \, s_{n-j} = - \sum_{n=0}^{L} s_{n-j} \, s_n \qquad (j = 1, 2, \ldots, k)$$

soit : $$\sum_{i=0}^{k} a_i \, v_{i-j} = - v_j \qquad (j = 1, 2, \ldots, k)$$

$$v_j = \sum_{n=0}^{N-1-|j|} s_n \, s_{n+|j|}$$

On aboutit donc bien dans ce cas aux mêmes équations auxquelles ont abouti ATAL-HANAUER et SAITO-ITAKURA. Pour la résolution de ce système MARKEL a utilisé la méthode du § 6.3, c'est-à-dire une méthode récurrente basée sur la définition d'un produit scalaire (ROBINSON).

6.4 Implantation asynchrone des méthodes d'identification. Mesure du fondamental

Pour une implantation synchrone, il faut connaître exactement l'instant de l'arrivée de la nouvelle période fondamentale. Ceci n'étant pas toujours facile, une implantation asynchrone des méthodes d'identification a été en général réalisée.

Un segment comprenant plusieurs périodes du pitch est considéré comme segment de traitement. Une fenêtre de Hamming est utilisée pour minimiser les effets de bord (Voir Fig 6.6). Pourtant, dans ce cas, il y a certaines régions (régions hachurées) où les suppositions déjà faites ne sont pas valables (le modèle recurrent du Chapitre 5 n'est pas valable). En considérant le segment de traitement comme un processus stationnaire on recherche un modèle pour ce segment. Le modèle est calculé par une des méthodes qui ont déjà été exposées. Notre but ici est de voir comment une mesure du pitch peut être effectuée sur le segment de traitement. Après avoir calculé un modèle la différence $\varepsilon_n = s_n - \hat{s}_n$ est effectuée pour tous les échantillons du segment. Cette différence donne une approximation de l'entrée. On voit des pics correspondant à cette entrée. Pourtant la détection de ces pics est aussi difficile que la détection des pics sur le signal vocal lui-même (voir figure 6.7).

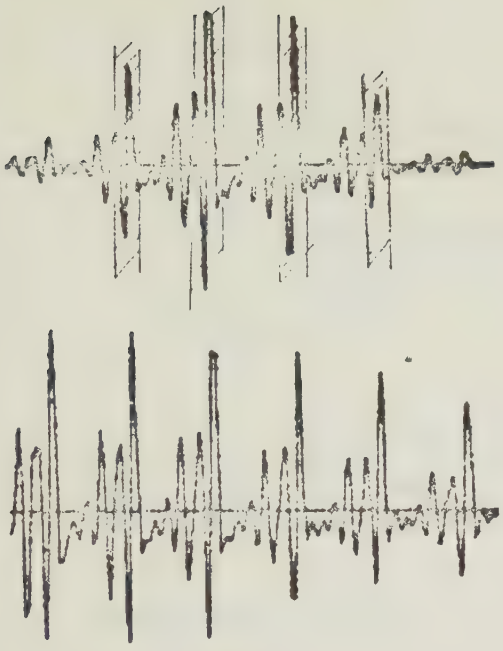


Figure 6.6 : Signal vocal vu sous une fenêtre de Hamming

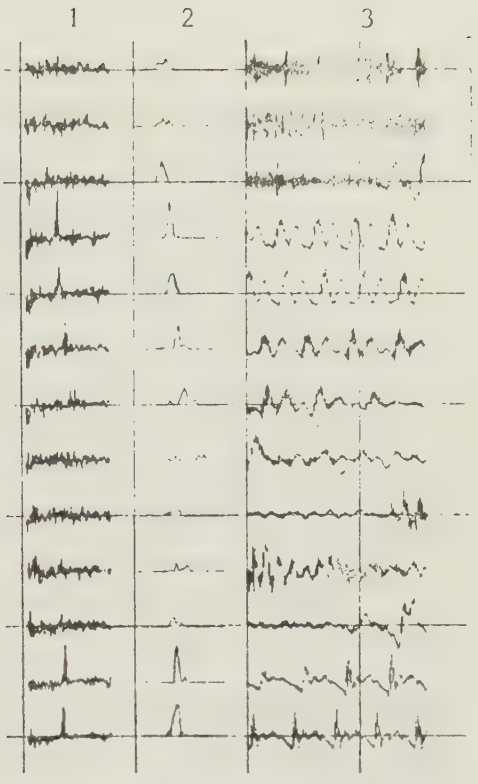


Figure 6.8 : Comparaison d'analyse des différents segments entre la méthode du Cepstre et celle d'autocorrélation d'un modèle de l'entrée (figure tirée de /14/) Sur la figure : 1) Méthode du Cepstre
2) Fonction d'Autocorrélation de l'entrée
3) Signal vocal

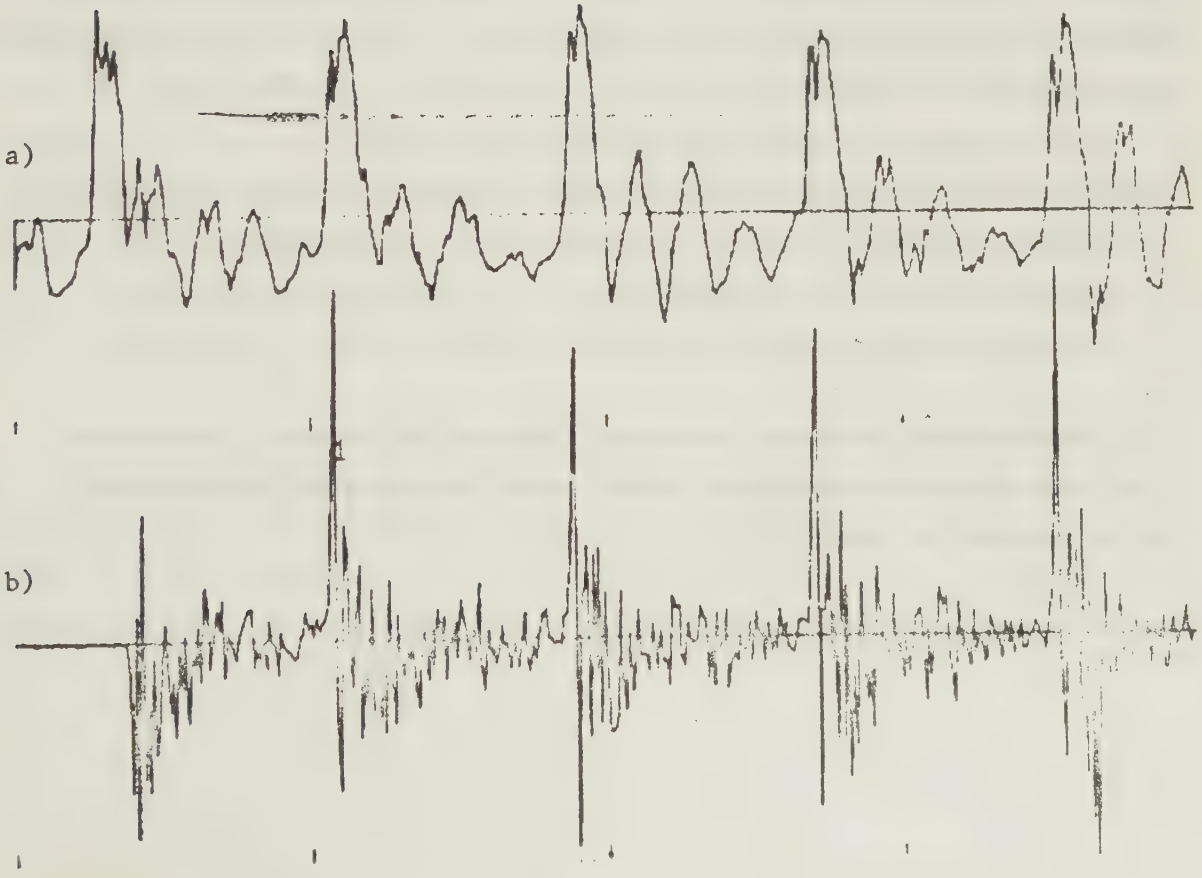


Figure 6.7 : a) Signal vocal b) Erreur de prédiction $\varepsilon = (s-\hat{s})$ (après identification

La solution pour calculer le pitch consiste à calculer la fonction d'autocorrélation de la séquence ε_n. Cette fonction est définie comme

$$r_n = \begin{cases} \sum_{j=0}^{N-1-n} s_j \, s_{j+n} , & n = 0,1,.....2N/2-1 \\ 0 , & n = 2\,N\,/\,2 \\ r_{2N-n} & n = 2N/2+1 \;,\; 2N/2+2, \;\ldots\; 2N-1 \end{cases}$$

Cette fonction présente un pic pour $t_n = T_o$, la période fondamentale. Sur la figure (6.8), on présente des résultats d'analyse de différents segments tels qu'ils ont été obtenus par MARKEL /14/. Ces résultats sont comparés avec des résultats issus de la méthode de cepstre. MARKEL fait la critique suivante de cette méthode.
Concernant la décision son voisé - non voisé, un simple test oui-non avec un seuil n'est pas suffisant pour déterminer si un segment est voisé ou non voisé. Des procédures plus compliquées sont donc envisagées.

Concernant la détermination du voisement sur une transition d'un son sonore à un son sourd, le problème n'est pas complètement résolu. Les mêmes problèmes se posent pour la méthode du cepstre. Les avantages de cette méthode par rapport à celle du cepstre sont les suivants:

1 - La décision voisé-non voisé est ici plus simple
2 - Le niveau dynamique du pic caractéristique du voisement dépend du nombre des périodes sous la fenêtre mais il ne dépend pas de la forme du spectre comme dans le cas de la méthode du cepstre.
3 - Une normalisation est possible par rapport au pic à l'origine.

En faisant diminuer l'ordre du système, le nombre d'opérations à effectuer diminue considérablement ce qui permet une implantation de la méthode en temps réel.

7. ANALYSE DE LA PAROLE PAR IDENTIFICATION RECURRENTE

7.1 Généralités

Le filtre de Kalman a été utilisé dans le but d'identifier une fonction de transfert approximant le comportement du système de phonation /33/, /34/, /35/, /36/. Une fonction en z (signal vocal intrinsèquement discret) a été utilisée. Un modèle contenant uniquement des pôles est adopté (voir § 5.1).

Cette méthode ressemble donc aux méthodes exposées précédemment dans la mesure où une fonction de transfert du même type est recherchée pour identification. Pourtant tout en leur ressemblant, elle dépasse beaucoup les autres méthodes grâce à des avantages théoriques ainsi qu'à des avantages de mise en oeuvre pratique.

En effet du point de vue optimisation, tout en étant une méthode de moindres carrés généralisés, elle sort beaucoup de ce cadre. Ceci parce qu'en même temps elle fournit une estimation au sens du maximum de vraisemblance (asymptotiquement optimale) ainsi qu'une estimation au sens du minimum de variance des paramètres à estimer. En outre, la possibilité de prise en compte des informations a priori concernant le processus implique une approche différente, inaccessible dans le contexte classique des moindres carrés. Par contre, dans le cadre de la méthode exposée ici cette introduction d'informations a priori peut se faire d'une manière tout à fait naturelle par l'intermédiaire des conditions initiales.

Du point de vue pratique une estimation en ligne est assurée (au fur et à mesure que les données arrivent). Par conséquent, une nouvelle estimation étant disponible avec un nouvel échantillon, la méthode est beaucoup plus informative. En outre, son caractère récurrent donne la possibilité d'implantation d'un algorithme synchrone de traitement

avec détection simultanée du pitch. Enfin la propriété de convergence rapide donne à l'algorithme des possibilités d' examen minutieux du signal. Nous allons donner dans la suite quelques notions théoriques sur le filtre de Kalman . Cette théorie ne sera pas donnée de façon rigoureuse . On tâchera juste de dégager les aspects " physiques " des équations d'estimation , ce qui nous permettra de les appliquer à notre problème particulier.

7.2. Le Filtre de Kalman digital

Nous resterons toujours dans le contexte digital . Ce filtre concerne l'estimation du vecteur d'état d'un système linéaire représenté par ses équations d'état .

$$\underline{x}_k = \Phi_{k,k-1}\, \underline{x}_{k-1} + F_{k-1}\, \underline{u}_{k-1} + \underline{w}_{k-1} \quad \text{équations d'état} \quad (7.1.)$$

x_k : est le vecteur d'état du système à l'instant k

$\Phi_{k,k-1}$: est la matrice de transition

F_k : est la matrice de commande

$\underline{w}_{k-1}$: est un vecteur de perturbation ayant des propriétés statistiques supposées connues et qui sont caractérisées par :

$$E(\underline{w}_k) = \underline{0} \qquad \text{(valeur moyenne nulle)} \qquad (7.2.)$$

$$E(\underline{w}_i\, \underline{w}_j^T) = \delta_{ij}\, Q_i \qquad \text{(perturbations indépendantes)}$$

Par ailleurs on suppose que le système est soumis à un certain nombre de mesures . Ces mesures peuvent être présentées comme les composantes d'un vecteur $\underline{z}_k$ donné par la relation suivante :

$$\underline{z}_k = H_k \underline{x}_k + \underline{v}_k \qquad \text{équations de mesure} \qquad (7.3.)$$

H_k est une matrice de sortie ou de mesure et v_k un vecteur " perturbation" ayant les caractéristiques suivantes :

$$E(\underline{v}_k) = \underline{0}$$
$$E(\underline{v}_i \, \underline{v}_j^T) = \delta_{ij} \, Q_i \qquad (7.4.)$$

En outre les suppositions supplémentaires suivantes seront faites

. indépendance des bruits du système et de mesure

$$E(\underline{w}_k \, \underline{v}_j^T) = 0 \qquad (7.5.)$$

. indépendance du bruit de mesure et de l'erreur initiale d'estimation

$$E(\underline{v}_n (\underline{x}_o - \hat{\underline{x}}_{o/o})^T) = 0 \qquad (7.6.)$$

A l'estimation du vecteur d'état $\hat{\underline{x}}$ on associe deux indices p et q ayant la signification suivante :

$\hat{\underline{x}}_{p/q}$ est l'estimation du vecteur d'état à l'instant t_p compte tenu des informations disponibles à l'instant t_q.

Par la suite une matrice de covariance $P_{p/q}$ sera attachée à $\underline{\hat{x}}_{p/q}$ cette matrice est définie comme :

$$P_{p/q} = E \left(\underline{x}_p - \underline{\hat{x}}_{p/q} \right) \left(\underline{x}_p - \underline{\hat{x}}_{p/q} \right)^T \qquad (7.6.)$$

L'équation d'état (7.1.) et l'équation de mesure (7.3.) donnent lieu à deux groupes d'équations d'estimation qui font l'objet de la théorie du filtre de Kalman .

$$\hat{x}_{k/k-1} = \Phi_{k,k-1} \, \underline{\hat{x}}_{k-1 \,/\, k-1} + F_{k-1} \underline{u}_{k-1} \qquad (7.7.)$$

équations de prédiction

$$P_{k/k-1} = \Phi_{k,k-1} \, P_{k-1/k-1} \, \Phi^T_{k,k-1} \qquad (7.8.)$$

$$\underline{\hat{x}}_{k/k} = \underline{\hat{x}}_{k/k-1} + K_k \left(\underline{z}_k - H_k \, \underline{\hat{x}}_{k/k-1} \right) \qquad (7.9.)$$

équations d'estimation

$$P_{k/k} = P_{k/k-1} - K_k \, H_k \, P_{k/k-1} \qquad (7.10.)$$

$$K_k = P_{k/k-1} \, H_k^T \left(H_k \, P_{k/k-1} \, H_k^T + R_k \right)^{-1} \qquad (7.11.)$$

Examinons ces équations un peu plus en détail .
L'équation (7.7) traduit les variations de l'état entre les instants t_{k-1} et t_k vu l' évolution du système . (le système n'étant pas stationnaire en général) . L'équation (7.9.) traduit la correction qu'on doit apporter au vecteur d'état $\underline{\hat{x}}_{k/k-1}$ compte tenu qu'à l'instant t_k on dispose de nouvelles informations (le résultat de mesure $\underline{z}_k$) .La matrice K_k permet le calcul d'un terme correctif en fonction de l'écart ($\underline{z}_k - H_k \, \underline{\hat{x}}_{k/k-1}$) et est appelée matrice de correction ou matrice de gain .

Le reste des équations (7.8), (7.10), (7.11) permet le calcul de cette matrice de correction d'une manière optimale .
La matrice K_k est calculée de façon à minimaliser la quantité :

$$\text{Trace } (P_{k/k}) = E \left(\underline{\hat{x}}_k - \underline{\hat{x}}_{k/k} \right)^T \left(\underline{x}_k - \underline{\hat{x}}_{k/k} \right)$$

donc de façon à avoir une variance d'erreur minimale . C'est dans ce sens que la méthode de Kalman est optimale . Sur ce point on va donner plus de précision par la suite . Le choix de la matrice de correction K_k peut, alors, être vueen relation avec la confiance attribuée à l'estimation précédente $\underline{\hat{x}}_{n/n-1}$, et en même temps avec l'intérêt porté aux nouvelles mesures . Il est à prévoir que plus la confiance dans l'estimation précédente sera faible , plus le gain K_k permettant de tenir compte de la nouvelle mesure devra être augmenté . En d'autres termes si la confiance en l'estimation précédente est déjà grande (c'est-à-dire trace $P_{k/k-1}$ petite) le filtre a tendance à considérer un éventuel écart valeur mesurée-valeur estimée, plutôt comme un bruit de mesure.

Application et analyse du signal vocal .

Le modèle du système de phonation est exprimé par l'équation

$$s_n + a_1 s_{n-1} + \ldots + a_k s_{n-k} = b_o e_n \qquad (7.12.)$$

avec la seule hypothèse que e_n est un bruit pseudo -blanc tel que

$$E(\underline{e}_n) = 0$$

$$E(\underline{e}_i \underline{e}_j) = 0 \quad \text{si } i \neq j \qquad E(\underline{e}_i \underline{e}_j) = \frac{\sigma^2}{b_o} \qquad (7.13.)$$

Cette hypothèse se justifie pour les sons sourds aussi bien que pour les

sons sonores où l'entrée de caractère impulsionnel peut être considérée de durée brève par rapport à la durée totale du processus . On peut par la suite reformuler l'équation(5.12.) sous la forme de variable d'état en écrivant :

$$\begin{aligned} \underline{a}_{n+1} &= I\,\underline{a}_n \\ \hat{\underline{s}}_n &= \underline{s}_n^T\,\underline{a}_{n-1} + v_n \end{aligned} \qquad (7.14.)$$

avec I matrice unité $k \times k$, $v_n = b_o\, e_n$ (7.15.)

$$a^T{}_n = a^T = \left[a_1 , a_2 , \ldots . a_k \right] \qquad (7.16.)$$

$$\underline{s}_n{}^T = \left[s_{n-1} , s_{n-2} , \ldots\ldots s_{n-k} \right] \qquad (7.17.)$$

Le problème est donc de déterminer une estimation optimale du vecteur d'état $\underline{a}_n = \underline{a}$ à partir de la mesure de la sortie s_n . Soit $\hat{\underline{a}}_p$ l'estimation à l'instant p et P_p la matrice de covariance correspondante.

$$P_p = E \left(\underline{a} - \hat{\underline{a}}_p \right) \left(\underline{a} - \hat{\underline{a}}_p \right)^T \qquad (7.18.)$$

On va adopter, conformément à la théorie précédente , comme critère d'optimalité le minimum de variance de l'estimation $\hat{a}_n$.
Soit $\varepsilon^2_n = \text{Trace} \left(P_n \right)$

Si on se restreint à un estimateur linéaire , on est conduit à définir la nouvelle estimation à l'instant n , résultant de la connaissance de s_n , par la relation de récurrence :

$$\hat{a}_n = \hat{a}_{n-1} + K_n \left(s_n - \underline{s}_n^T \hat{a}_{n-1} \right) \qquad (7.19.)$$

où le coefficient vectoriel K_n pondère les influences repectives de l'ancienne estimation et de la différence entre la sortie réelle s_n et la sortie attendue $\underline{s}_n^T \hat{\underline{a}}_{n-1}$

La détermination du gain optimal fait comme on l'a déjà vu l'objet du filtre de Kalman .

Les équations récurrentes définissant la solution se réduisent alors à l'ensemble :

$$K_n = P_{n-1} \underline{s}_n \left(\underline{s}_n^T P_{n-1} \underline{s}_n + \sigma^2 \right)^{-1} \qquad (7.20.)$$

$$P_n = \left(I - K_n s_n^T \right) P_{n-1} \left(I - K_n s_n^T \right)^T \qquad (7.21.)$$

ou encore

$$P_n = P_{n-1} - K_n s_n^T P_{n-1} \qquad (7.22.)$$

en utilisant cette dernière relation on peut montrer que la contribution de la mesure s_n se traduit par :

$$\text{Trace } (P_n) \leqslant \text{Trace } (P_{n-1}) \qquad (7.23.)$$

qui correspond à une amélioration de l'estimation à chaque fois .

On retrouve les relations précédentes à partir des formules générales données dans le paragraphe précédent en posant :

$$\underline{x}_k = \underline{a}$$

et en se limitant au cas stationnaire :

$\underline{x}_k = \underline{x}_{k-1} = a$, ce qui donne $\Phi_{k,k-1} = I$; $F_k = 0$; $Q_k = 0$

en outre : $H_k = \underline{s}_k^T = \left[s_k , s_{k-1} , \ldots , s_{k-m+1} \right]$ et $z_k = s_k$

Donc les formules sont valables dans le cas d'un signal stationnaire . dont l'entrée peut être considérée comme un bruit pseudo blanc .

Pour $Q_k = Q \neq 0$ la formule (7.22.) seule est différente .

On aura dans ce cas :

$$P_n = P_{n-1} - K_n \underline{s}_n^T P_{n-1} + Q \qquad (7.24.)$$

La matrice Q tient compte des propriétés de dispersion du vecteur d'état .

7.4. Mise en oeuvre de l'algorithme .

a) Initialisations

L'algorithme se prête à un grand nombre de variantes par le jeu des paramètres relatifs à l'initialisation . / 30 /

Les valeurs initiales $\hat{a}_o$ et P_o dans (7.19) et (7.20.) permettent d'intégrer par la méthode des informations à priori . On peut ainsi considérer $\underline{\hat{a}}_o$ et P_o comme une estimation initiale accompagnée de sa matrice de covariance .Dans le cas où l'on ne dispose d'aucune information à priori sur $\underline{a}$, nous pouvons commencer avec une valeur de $\underline{a}$ finie et quelconque . Concernant la matrice P_o , une matrice de covariance diagonale peut être choisie (ceci traduit notre incapacité d'introduction d'une correlation quelconque). Les termes diagonaux de cette matrice doivent être infinis , traduisant ainsi l'arbitraire de l'estimation initiale (variance infinie) Nous aurons par exemple :

$$\underline{a}_o = \underline{0} \quad ; \quad P_o = \lim_{a \to \infty} (a^2 I) \qquad (7.25.)$$

La matrice d'information correspondante sera alors $I_o = P_o^{-1} = 0$

Le fait de pouvoir disposer d'une initialisation plus réaliste du couple (a,P) pourrait contribuer à des réductions importantes du temps de calcul . Cette initialisation consiste en la détermination d'une certaine structure de la matrice P , liée à un certain nombre des connaissances a priori . Ces connaissances peuvent être , par exemple , une détermination de P à partir des expériences antérieures .

On a montré que la convergence est notablement accélérée par le choix d'une matrice P_o sophistiquée /30 /

$$P_o : \{ p_{ij} = \sigma_P^2 \exp (- \frac{1}{2} (\frac{i - j}{\theta})^2) \} \qquad (7.26.)$$

On peut aboutir à une telle stucture pour la matrice P_o , si on tient compte du fait que les coefficients a_i ne sont pas indépendants (hypothèse correspondant à l'absence totale de connaissances a priori) mais sont solution d'une équation récurrente étant donnée la relation existante entre a_i et s_i /30/ . La relation (7.26.) nous dit que deux coefficients a_i et a_j dont les indices i et j ne sont pas très différents se présenteront plus corrélés . Au contraire si l'écart des i et j est grand les coefficients a_i et a_j se présenteront moins corrélés . Le paramètre θ fixe le pourcentage de corrélation , plus θ est grand , moins les paramètres seront corrélés . Dans la pratique on peut prendre

$$k \leq \theta \leq \frac{k}{10} \qquad (k \text{ étant l'ordre du système })$$

Concernant le choix de σ_P^2 (7.26.) ou de a^2 (7.25) . Il faut noter que des valeurs trop importantes dans P_o sont en général progressivement corrigées par le filtre lui-même . Il faut noter que le choix des P_o faibles conduit à des filtres timides et que le choix des P_o trop élevés conduit à des filtres trop entreprenants . On evaluera les conséquences de ce choix par la suite.

b. Choix de σ^2

Jusqu'à présent la variance du bruit σ^2 a été supposée connue . Pourtant sur le signal vocal ce n'est pas le cas . Une estimation de σ^2 à partir du rapport signal / bruit , en outre , n'aurait pas de sens vu que dans notre cas σ^2 est plutôt lié aux caractéristiques de l'entrée . Une estimation de bruit pourrait être obtenue après identification /30 / .

$$\hat{\sigma}^2 = \frac{1}{N-k} \left(\underline{s} - S\underline{\hat{a}} \right)^T \left(\underline{s} - S\underline{\hat{a}} \right) \qquad (7.27.)$$

Le choix de $\underline{\sigma}^2$ est en relation avec la nervosité du filtre qui va résulter . Nous pouvons voir ceci à partir des relations (7.20.) et (7.22.) Ainsi

$$K_n = P_{n-1} \underline{s}_n \left(\underline{s}_n^T P_{n-1} \underline{s}_n + \sigma^2 \right)^{-1}$$

soit

$$K_n \left(\sigma^2 + \underline{s}_n^T P_{n-1} \underline{s}_n \right) = P_{n-1} \underline{s}_n$$

$$K_n \sigma^2 + K_n s_n^T P_{n-1} \underline{s}_n = P_{n-1} \underline{s}_n$$

$$K_n \sigma^2 + \left(- P_n + P_{n-1} \right) \underline{s}_n = P_{n-1} \underline{s}_n$$

$$K_n = \frac{1}{\sigma^2} P_n \underline{s}_n \qquad (7.28.)$$

Alors plus σ^2 est grand , moins le gain K_n a tendance à être petit , donc pour de grandes valeurs de σ^2 le filtre a tendance à être timide . Au contraire pour de petites valeurs de σ^2 le filtre est en général entreprenant .

En outre pour σ^2 donné ,(7.28) traduit la relation existant entre la nervosité du filtre en question et la valeur de la matrice de covariance P_n . Ainsi plus la matrice de covariance est importante plus le filtre est nerveux .

Dans la pratique l'identification s'effectue même si les valeurs de σ^2 et σ_p^2 ne sont pas les bonnes . Ces filtres sont en général qualifiés de sous-optimaux . Si l'estimation des variables d'état n'est pas perturbée, ce n'est pas le cas pour les estimations de P_n qui sont en général fausses .

c. Automatisation de l'algorithme.

Le système de phonation variant dans le temps , il convient d'initialiser périodiquement l'algorithme pour actualiser le modèle . Cette opération peut , en particulier , être déclenchée automatiquement en synchronisme avec les périodes fondamentales . Un décalage systématique peut être introduit pour n'appliquer l'algorithme que sur des segments où l'excitation peut être considérée de moyenne nulle . Pour les sons non voisés , une durée arbitraire de quelques ms est retenue. Une mise en oeuvre asynchrone a été réalisée, par ailleurs , pondérant le signal par une fenêtre glissante .

Un exemple de mise en oeuvre est donné par la figure (7.1.) pour une fréquence d'échantillonnage de 10 kHz et un ordre k = 12 . La convergence de l'algorithme est , en moyenne , réalisée après 30 échantillons (soit 3 ms , durée inférieure à la période fondamentale).

7.5. Applications pratiques du modèle.

a. Détection du fondamental.

La détection des périodes fondamentales peut être réalisée par l'évolution après convergence des grandeurs :

$$\varepsilon_n = s_n - \hat{s}_n \quad , \quad \delta_n = \underline{\hat{a}}_n \underline{\hat{a}}_n^T - \underline{\hat{a}}_{n-1} \underline{\hat{a}}_{n-1}^T \qquad (7.29.)$$

$$\gamma_n = (\underline{\hat{a}}_n - \underline{\hat{a}}_{n-1}) (\underline{\hat{a}}_n - \underline{\hat{a}}_{n-1})^T , \quad \beta_n = \frac{1}{\text{trace}(P_n)} \frac{d\,(\text{trace}\,(P_n))}{dt}$$

Les critères intrinsèques δ_n et γ_n , sont accessibles seulement par une méthode récurrente . Le critère β_n est particulier à l'utilisation du filtre de Kalman pour estimer K_n . Nous avons surtout utilisé le critère γ_n . Une estimation $\hat{a}_n$ du système étant disponible à chaque instant , une détection en ligne du pitch est possible . Les remarques qui suivent portent sur le critère γ_n en comparaison avec le critère ε_n qui a été utilisé par ATAL /10/ et MAKSYM /37/ .
Le critère γ_n n'est pas très sensible aux bruits . En effet après convergence les différences entre sortie réelle et sortie estimée peuvent plutôt être attribuées aux bruits (système stationnaire) ainsi ces différences ne se traduisent pas par des variations du système . En outre quand une nouvelle impulsion du pitch arrive la différence (s-ŝ) reste très importante pour que le filtre la considère comme un bruit de mesure . Ainsi des variations importantes se manifestent au système . De cette façon nous pouvons imaginer la fonction $\gamma_n = \gamma_n (s - \hat{s})$ comme une fonction de gain sélectif . Ainsi une action amplificatrice s'exerce aux différences (s-ŝ)qui sont dûes à des variations importantes du système . Par ailleurs ces différences sont réprimées dans le critère γ_n si elles sont dues à des bruits de mesure . Sur la figure (7.1.b) on voit comment la détection du pitch est effectuée.

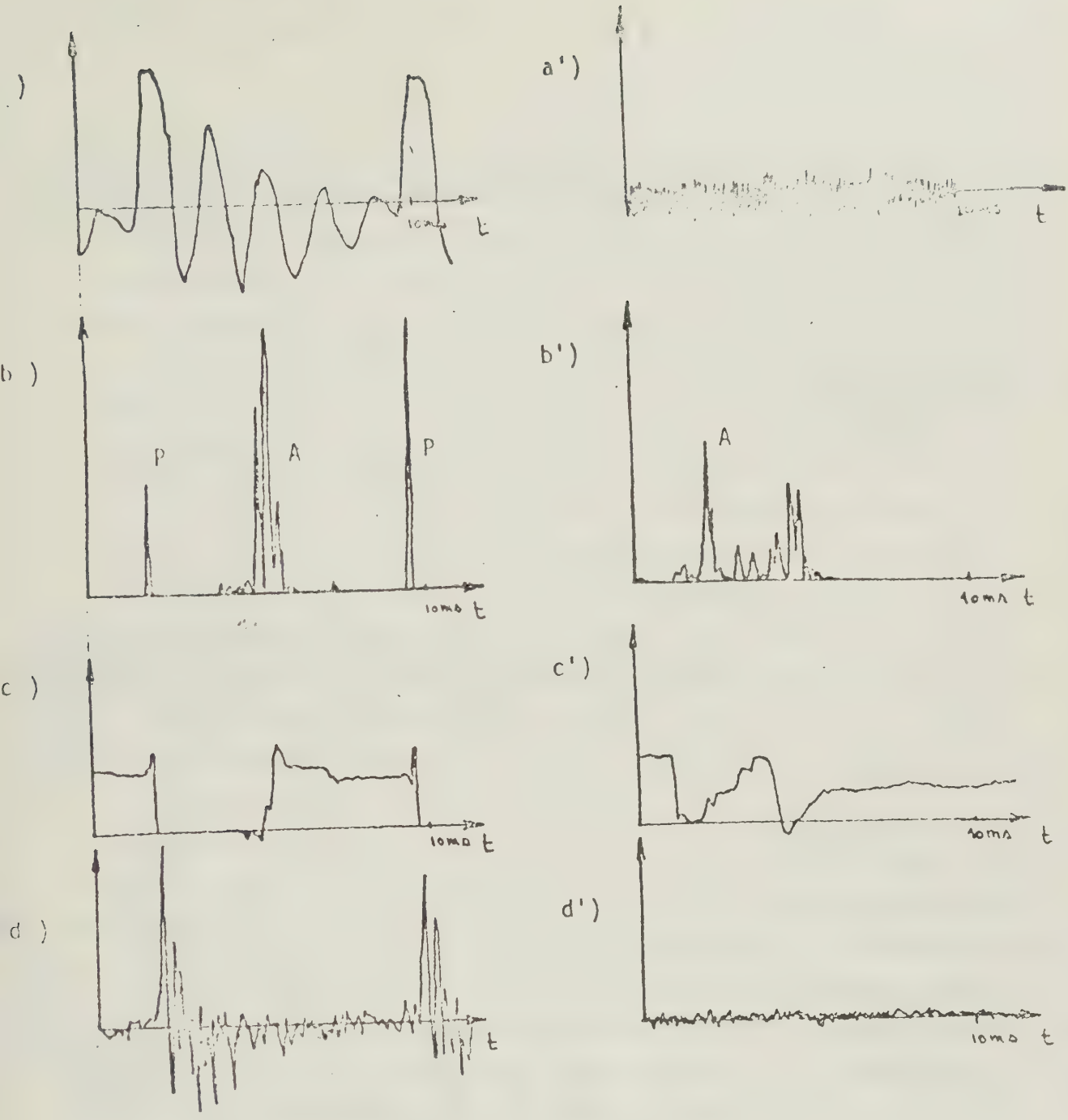


FIGURE 7.1 : Mise en oeuvre de l'algorithme.

a) Son sonore a

b) Variation du critère γ_n.

c) Evolution du parametre a_3.

d) Entrée calculée .

a') Son sourd s

b') Variation du critère γ_n.

c') Évolution du paramètre a

d') Entrée calculée

Pics P : détection du pitch

Pics A : initialisation de l'algorithme .

Nous avons vu au § 7.4 comment avec les valeurs (σ^2, σ_p^2 et θ) on pouvait conditionner la nervosité du filtre . Ainsi si notre intérêt principal n'est pas l'estimation des paramètres du système mais la détection du pitch, nous avons remarqué qu'un tel conditionnement était possible pour que le filtre soit particulièrement sensible en vue de cette détection (par exemple, filtre trop timide ou trop entreprenant). En outre, un ordre de système k = 2 ou 3 (qui a l'avantage supplémentaire d'assurer une variabilité réduite du modèle), peut être suffisant pour une telle détection. On peut ainsi, arriver à travailler en temps réel. En effet, le nombre de multiplications à effectuer étant de l'ordre de $3k^2/2$ par itération pour que k = 3, on aura 14 multiplications/itération. Pour une fréquence d'échantillonnage f = 10 kHz, nous avons 100 µsec pour effectuer ces opérations. La figure 7.2 nous montre la détection du pitch sur une partie d'une phrase. La figure 7.3 nous montre les variations du logarithme du critère γ_n sur une partie d'une autre phrase. Deux cas différents sont présentés : a) Fonctionnement sans réinitialisation après chaque détection de pitch; b) Fonctionnement avec réinitialisation. Nous remarquons qu'à partir d'un certain nombre d'itérations le filtre a tendance à perdre se nervosité. Il est donc de plus en plus difficile de détecter le pitch. Pourtant, il faut ajouter qu'avec un filtre bien conditionné pour détecter le pitch on peut se permettre de réinitialiser à des intervalles assez éloignés (de l'ordre de 1 seconde). Les performances de la méthode sont très bonnes, mais il faut noter ses difficultés à détecter le pitch sur les transitions d'un son sonore vers un son sourd. En effet, sur les transitions le modèle utilisé peut être mis en cause. Il n'y a pas à notre connaissance d'algorithme qui détecte le pitch dans le cas mentionné.

b. Calcul des pôles, du spectre et des formants à partir des coefficients du modèle

Nous allons rechercher par la suite comment notre modèle peut se lier d'une manière plus directe et plus naturelle avec le système de phonation, par exemple, avec des grandeurs comme les fréquences

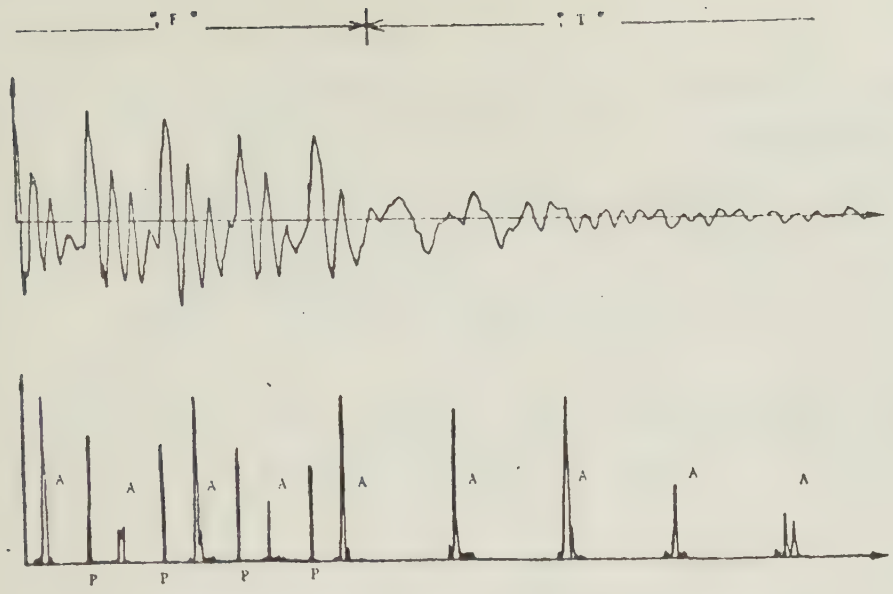
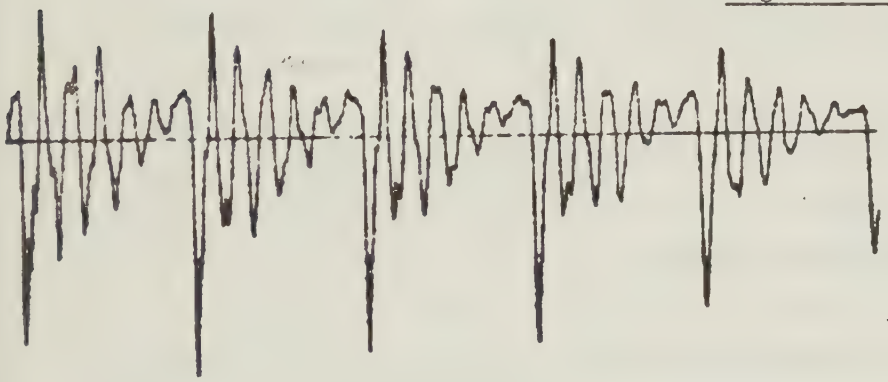
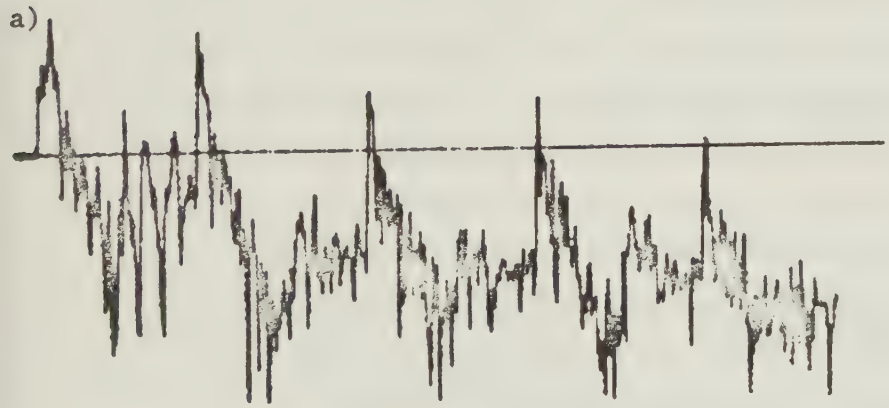


Figure 7.2 : Détection synchrone des débuts de période fondamentale.
Pics A : initialisation de l'algorithme.
Pics P : début de période fondamentale (pitch).

Signal vocal





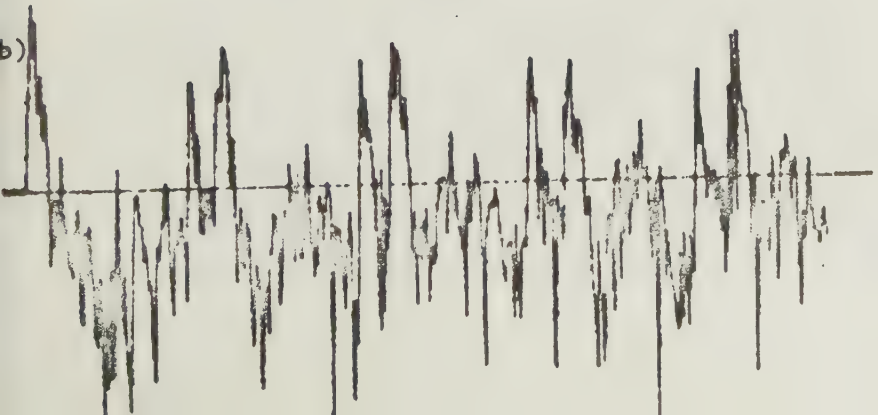


Figure 7.3 : Log(γ_n) en fonction du temps.
a) Fonctionnement de l'algorithme sans réinitialisation
b) Fonctionement avec réinitialisation après chaque détection du pitch.
(σ_p=10 , θ=2 , σ^2=0.9)

naturelles de résonance (formants) ou le spectre en fréquence etc . Avant d'aborder ce sujet nous discuterons ce qui concerne les pôles de la fonction de transfert en z . Les pôles de la fonction en z sont les racines de l'équation :

$$1 + \sum_{i=1}^{k} a_i \, z^{-i} = 0 \qquad (7.30.)$$

Pour assurer la stabilité nous savons que ces pôles doivent être à l'intérieur du cercle unitaire sur le plan des z Ceci n'est pas toujours le cas , et dans la pratique à cause des différentes erreurs d'arrondi nous arrivons à trouver des instabilités , bien que théoriquement le filtre récursive du type utilisé (mémoire finie) soit en général stable . Ces instabilités sont donc purement numériques et liées à la compétence de la méthode de calcul .

Comme on l'a vu au chapitre 6 différentes approches et différents algorithmes ont été utilisés dans le but d'ajustement d'un modèle tel que le modèle (5.1.) . Parmi les différents chercheurs ATAL-HANAUER avec une approche des moindres carrés et en utilisant la méthode de Cholensky comme algorithme de résolution (voir Chapitre 6) constatent le besoin d'utilisation d'une correction a posteriori du modèle pour se ramener à un modèle stable . La correction consiste a recaler les pôles quand ils sont sortis du cercle unitaire . Ensuite les paramètres sont recalculés à partir des pôles corrigés . . Mais cette correction demande chaque fois le calcul des pôles ce qui est une opération très coûteuse en temps de calcul. ATAL - HANAUER ne donnent pas la fréquence d'apparition des pôles instables ni les raisons . En outre les conditions de calcul ne sont pas données non plus . SAITO - ITAKURA , ont utilisé le même modèle avec une approche et un algorithme de résolution différents . Pourtant la stabilité n'est pas discutée . Il est dit "qu'un modèle stable est obtenu " . Plus tard MARKEL /29/ utilisant le même algorithme de résolution que SAITO-ITAKURA a examiné à fond le problème de stabilité

dans le cas de cet algorithme de résolution (Méthode de Robinson)
(voir Chapitre 6)
MARKEL a abouti aux résultats suivants :

- des modèles instables sont obtenus parfois à cause d'accumulations des erreurs d'arrondi pendant le calcul .

- La méthode de calcul est à l'origine de ces erreurs.

- La fréquence d'apparition des modèles instables dépend

    1. De l'ordre du modèle
    2. De la largeur du mot mémoire utilisé pendant le calcul

- Cette fréquence est très élevée pour les sons sonores en comparaison avec celle relative aux sons sourds.

En général les sons sourds fournissent un modèle stable . Ceci est expliqué par MARKEL à partir des différences au niveau dynamique . Entre les sons voisés et les sons non-voisés , ce qui fait que beaucoup plus d'erreurs d'arrondi sont cumulées pour les sons voisés à cause de leur plus forte dynamique . MARKEL a tracé des courbes très intéressantes qui donnent la probabilité de modèles instables en fonction de l'ordre du système et de la longueur du mot mémoire . (voir fig 7.4.)

On s'apperçoit qu'à partir de longueurs $>$ 32 bits cette probabilité commence à devenir minime . Ceci est une chose assez inquiétante vu la relation du prix de revient d'un appareil spécialisé en hardware et la longueur du mot mémoire utilisée pour les calculs .

Concernant la méthode basée sur le filtre de Kalman , jusqu'à présent nous n'avons pas fait de recherche systématique sur la stabilité . Nous avons constaté l'influence de l'ordre du système et de la longueur du mot mémoire et nos conclusions ressemblent à celles de MARKEL . Pourtant nous avons observé les points suivants .

La méthode utilisée est moins sensible aux erreurs d'arrondi étant donné l'utilisation de l'estimation $\hat{a}_k$ comme variable participant aux calculs (caractère adaptatif de l'algorithme)

Etant donné le caractère récurrent de la méthode on peut exactement affirmer à quel moment l'instabilité se déclare . De cette façon on pourra chercher des raisons d'instabilité liées plutôt à la nature du problème .

En accord avec ces points , nous avons constaté les choses suivantes :

1. Les modèles instables étaient plus fréquents sur les sons sonores et moins fréquents sur les sons sourds . Ceci nous paraît être lié à l'influence de l'entrée (la relation de prédiction $s_n = \sum_{i=1}^{k} s_{n-i} a_i$ n'étant pas valable quand l'entrée apparaît).

Nous avons expérimenté ce point de vue . Nous avons initialisé l'algorithme au point A . (voir fig 7.5.) un modèle est fourni à tout instant. Le modèle était stable sauf dans la région (C) fortement influencée par l'entrée . On donne ces résultats sur la figure 7.6.

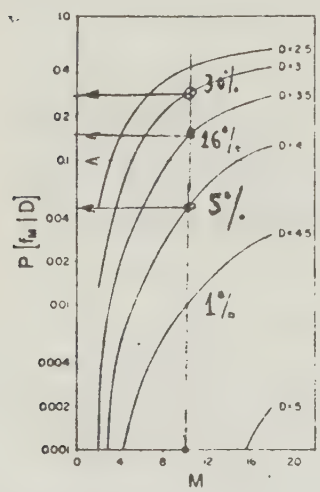


Figure 7.4 : Probabilité de modèle instable en fonction de l'ordre du système et de la longeur du mot mémoire. (Figure tirée de /29/).

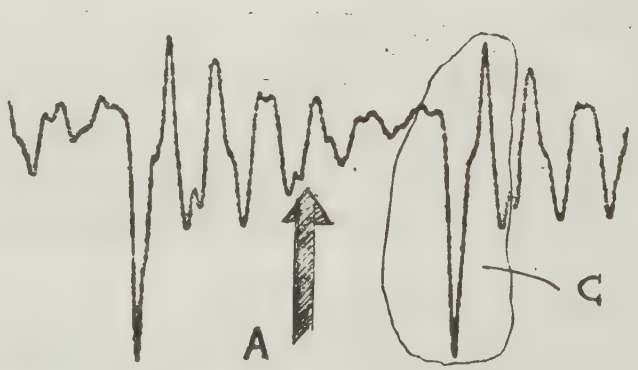


Figure 7.5 : Initialisation de l'algorithme au point A. Passage par la région C.

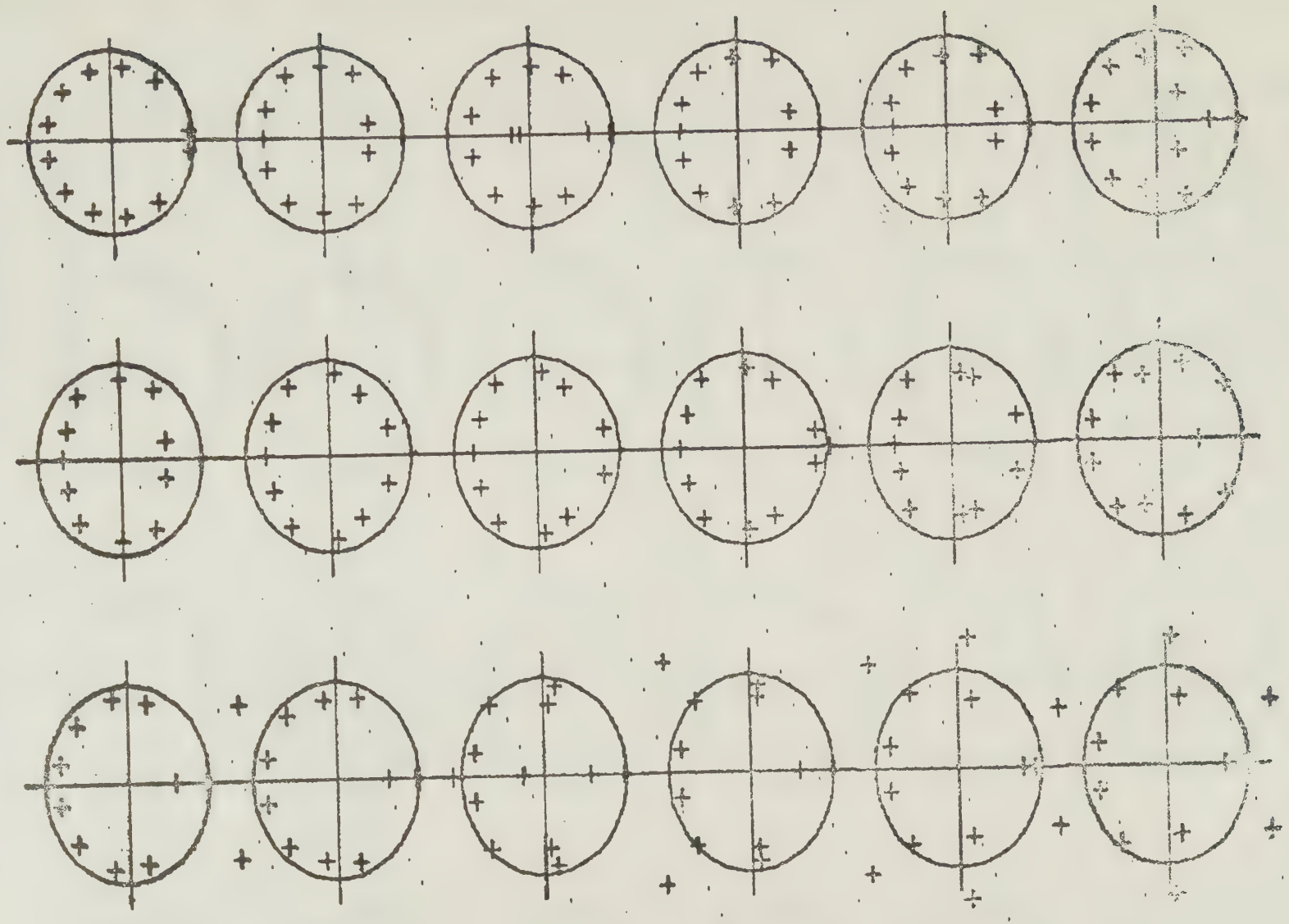
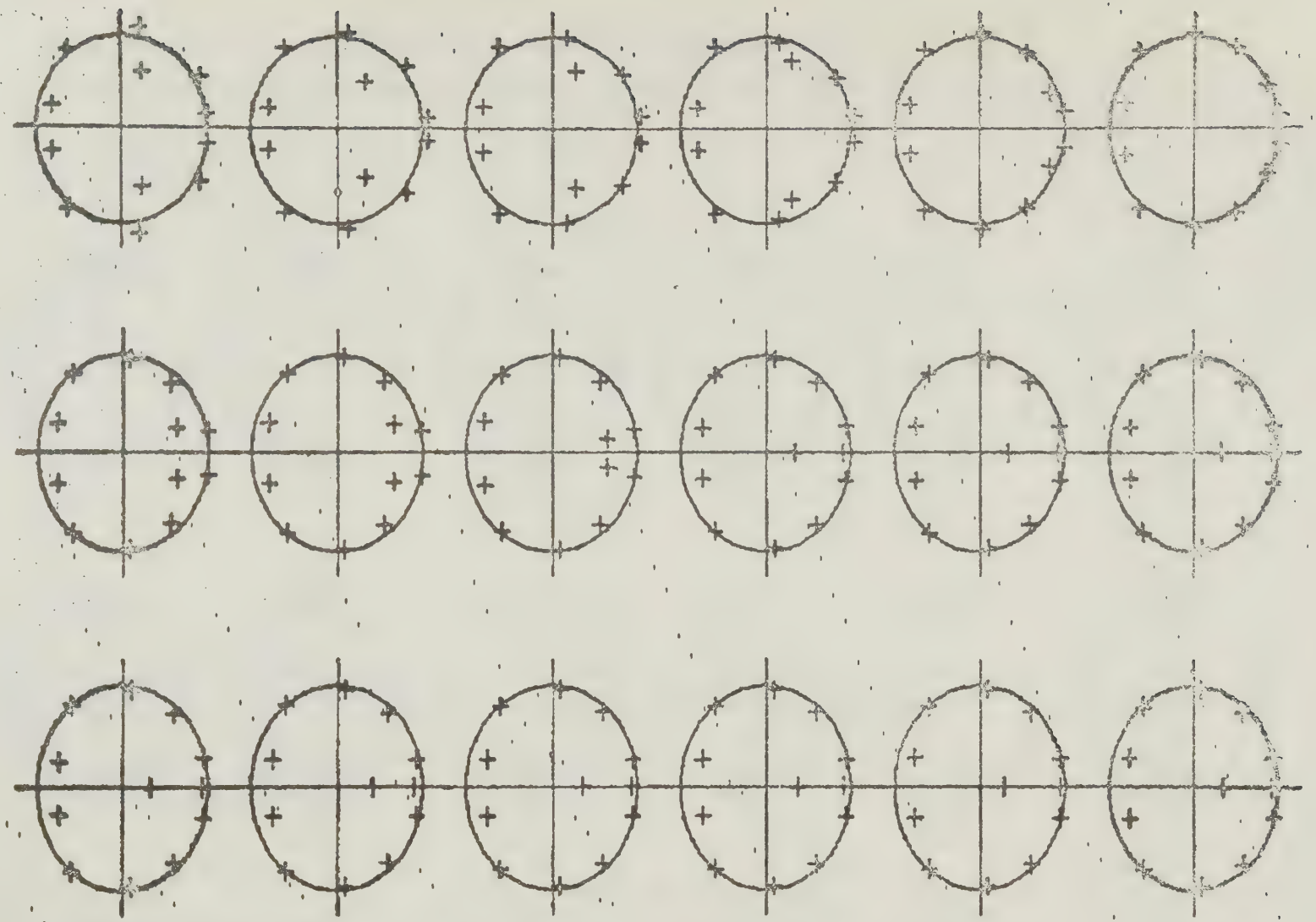


Figure 7.6 : Experience pour demontrer l'instabilitée créée par le pitch quand il apparait à un instant où l' algorithme n'a pas encore bien convergé.



Suite de la figure 7.6 : Après le passage par la région du pitch le système devient de nouveau stable

2. Les modèles instables étaient plus fréquents sur les transitions , en particulier sur celles d'un son sourd vers un son sonore (faible dynamique ⟶ forte dynamique) .
(représentativité du modèle mise en question)
En ce qui concerne les racine de l'équation (7.30.) des algorithmes basé sur la méthode de Newton - Raphson ont été trouvés assez efficaces .
Calcul du spectre : Une approximation de l'en eloppe du spectre peut être obtenue à partir de la fonction de transfert H (z)
Cette enveloppe sera donnée par la relation :

$$\left| H (\omega) \right|^2 = \frac{b_o^2}{\left| 1 - \sum_{m=1}^{k} a_m e^{-jm \omega T} \right|^2} = \left| H (z) \right|^2_{z=\exp (j\omega T)} \qquad (7.31.)$$

Le calcul de ce spectre peut être effectué facilement si l'on utilise la transformée de Fourier rapide (F.F.T.) à partir de la formule suivante

$$H (\omega) = \frac{b_o^2}{\text{FFT} \ (1 , -a_1 , -a_2 , \ldots , -a_k , 0 , \ldots , 0)} \qquad (7.32.)$$

Les zéros sont ajoutés pour augmenter la résolution du spectre obtenu .
Sur la figure (7. 7) on voit des spectres calculés de cette façon .
Nous constatons des pics très aigus . Ceci est spécifique de méthodes de modèles en général et correspond à des pôles très près de la circonférence du cercle unitaire .

Calcul des formants :

Deux méthodes existent pour calculer les formants à partir des coefficients

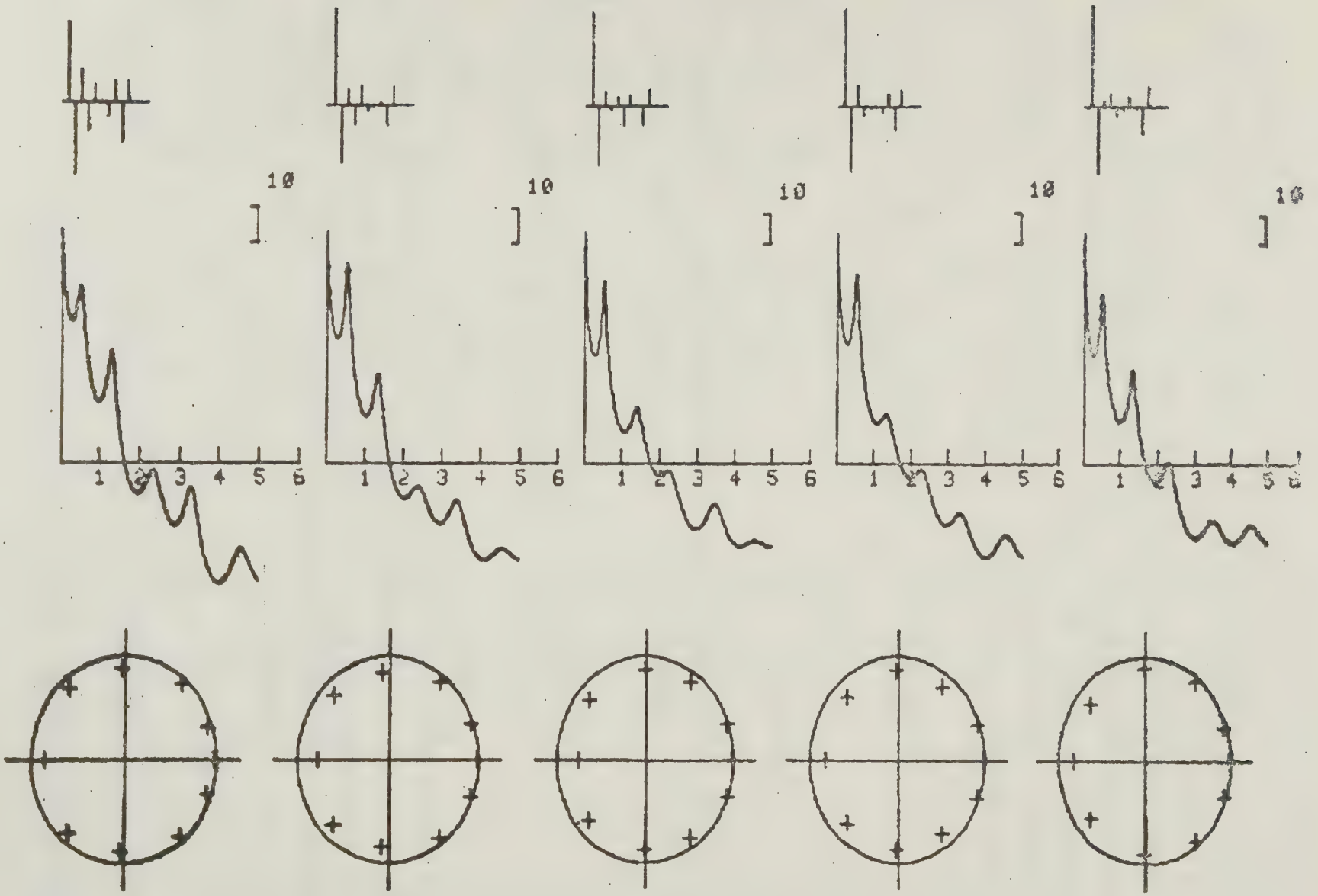


Figure 7.7 : Evolution des spectres et pôles sur la région stable d'un phonème (/a/).
Analyse effectuée avec le filtre de Kalman.

de prédiction . La première consiste à calculer le spectre à partir de la formule 7.32 et à appliquer par la suite sur ce spectre un algorithme de détection de pics.

La seconde méthode consiste à calculer les pôles et par la suite les formants à partir de ces pôles. Si z_i est un pôle sur le plan des z , la fréquence et l'amortissement correspondants peuvent être calculés numériquement à partir des formules :

$$F_k = \frac{1}{2\pi T} \text{Im} (\ln z_k)$$

$$B_k = \frac{1}{\pi T} \text{Re} (1/ \ln z_k) \qquad (7.33)$$

7.6 Discussion

L'estimation récurrente a été prouvée très bien adaptée au traitement du signal vocal. Pourtant nous avons vu qu'un nombre d'opérations très important étaient nécessaires pour le calcul du gain optimal par la méthode de Kalman . En outre l'optimalité de la méthode peut être discutable étant donné que les caractéristiques des bruits ne sont pas connues. Nous avons déjà dit que le filtre de Kalman était une des méthodes possibles pour estimer le gain K_n de l'equation :

$$\underline{a}_{n+1} = \underline{a}_n + K_n (s_n - \hat{s}_n) \qquad (7.34)$$

Pourtant un éventail d'autres méthodes existent pour l'estimation de K_n . Ces méthodes sont en général qualifiées de sous optimales mais elles conduisent parfois à de bons résultats. (Voir /38/ , /39/ , /40/). Le vocable " Méthodes d'identification stochastique " (stochastic identification) ou encore " Identification par apprentissage " (learning identification) est souvent utilisé pour désigner ces méthodes . Leur qualité peut être jugée à la base d'une courbe qui exprime les propriétés de convergence et qui est caracté-

ristique des conditions de convergence de chaque méthode. Nous donnons cette courbe sur la figure (7.8). Il s'agit de la variation d'une mesure de distances (ordonnées) avec le temps (abscisses). La distance peut être, par exemple, une distance de structure :

$$d^2 \triangleq \left|\underline{\hat{a}}^j - \underline{a}\right|^2 = \sum_{i=1}^{k} \left(\hat{a}_i^j - a_i \right)^2 \qquad (7.35)$$

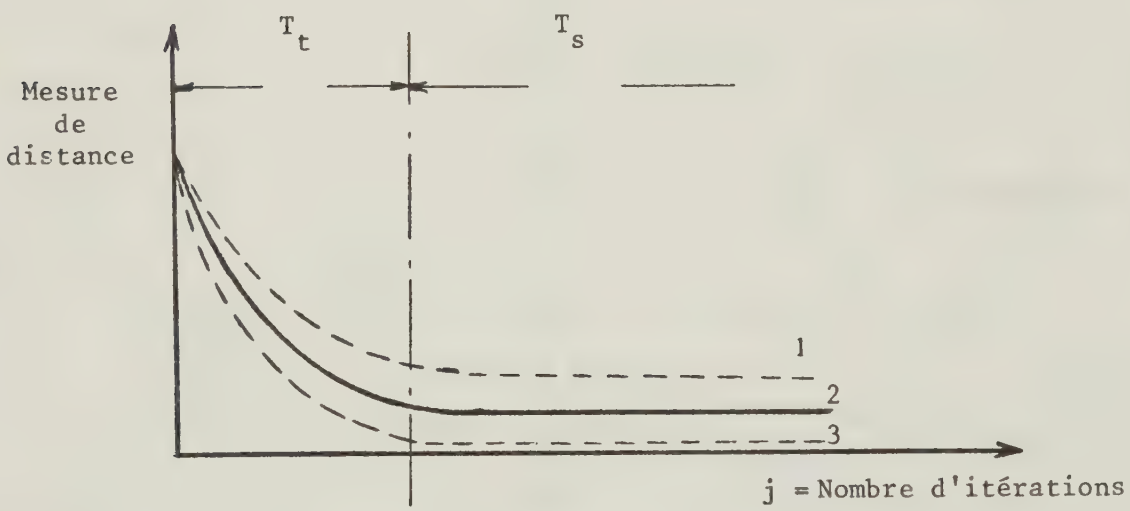


Figure 7.8 : Période "transitoire" et période "stable" d'un algorithme récurrent

Sur la figure deux durées différentes sont déterminées :

a. La durée transitoire (T_t) : représente le temps que l'algorithme met pour converger. Cette durée est en général variable d'un algorithme à l'autre. La rapidité de convergence est une qualité qui est souvent demandée (enregistrement de longueur limitée, stationnarité à court terme, etc ...). Parmi les algorithmes dont on dispose actuellement, c'est le filtre de KALMAN disposant des informations a priori qui donne le T_t minimum.

b. La durée stable (T_s) : il s'agit du temps après convergence. La caractéristique essentielle de cette durée est la fluctuation des paramètres autour d'une valeur. Cette fluctuation est minimale et la méthode du filtre de Kalman est utilisée.

On étudie actuellement différentes méthodes d'approche sous-optimale. Cette étude est facilitée vu l'implantation modulaire de l'algorithme récurrent déjà étudié (filtre de Kalman). Ainsi, pour changer de méthode, il suffit de changer un sous-programme. Le but de l'étude est de voir si certaines méthodes sous-optimales pouvaient s'adapter au traitement du signal vocal ce qui aurait comme conséquence des temps de calcul réduits (nous avons vu que dans le cadre de plusieurs applications la notion du temps réel était importante pour un algorithme d'analyse du signal vocal, voir chapitre 3).

L'équation (7.34) peut s'écrire :

$$\underline{a}_{n+1} = \underline{a}_n + B_n \underline{s}_n \ (s_n - \hat{s}_n)$$

Les approches d'étude se situent entre la méthode de gradient

$$(B_n = b_n . I \ , \ b_n = \frac{1}{n+1} \ \frac{1}{\left| \underline{s}_n \right|^2})$$

et le filtre de Kalman

$$B_n = P_{n-1} \ (\sigma^2 + \underline{s}_n^T \ P_{n-1} \ \underline{s}_n)^{-1}$$

Au stade actuel, on ne peut rien affirmer sur la qualité des différentes méthodes. Nous avons constaté que les valeurs exactes des variances des bruits impliqués étaient en général nécessaires pour la convergence. Ceci semble être l'inconvénient de ces algorithmes en comparaison avec les approches basées sur un critère de moindres carrés.

7.7. Réalisations , programmes .

Dans le cadre de cette étude nous avons envisagé un certain nombre de buts .

1. Mise en oeuvre d'une nouvelle méthode (filtre de Kalman)
2 Possibilité de comparaison facile avec les autres méthodes utilisées dans le cadre de l'analyse de la parole .
3. Facilitées en vu d'une recherche fondamentale sur n'importe quel signal.
4. Modularité
5. Préparation de la recherche dans l'avenir

Les programmes établis répondent à ces buts . Ainsi la mise en oeuvre de la méthode basée sur le filtre de Kalman étant notre but principal , nous avons envisagé cette inovation dans le cadre de tout ce qui a préexisté . Ceci pour rendre toute comparaison possible . Des programmes conversationnels avec plusieurs options ont eu comme but non seulement de tester les méthodes dont ils faisaient l'objet , mais aussi de constituer un outil de recherche fondamentale dans le cadre de l'analyse du signal . Ainsi ces programme n'étant pas établis à la base du signal vocal peuvent avoir comme entrée n'inporte quel autre signal à traiter (houle , électrocardiogramme , encéphalogrammes; etc) qui pourrait être approximé par un modèle analogue à celui du chapitre 5 .

quatre catégories principales d'algorithmes ont été établis

a. Algorithmes conversationnels d'analyse récurrente
b. " " " globale
c. " automatiques " et de synthèse récurrente
d. " " " et de synthèse globale .

Avec cet éventail d'algorithmes, nous pouvons couvrir la plupart des méthodes d'analyse numérique qui existent actuellement dans le cadre de l'analyse de la parole, ainsi que la simulation des vocoders correspondants. De cette façon une autre méthode peut être utilisée à chaque instant par l'appel d'un sous-programme différent dans le contexte (a) et (c) ou (b) et (d) (modularité).

En outre le dépouillement des résultats de ces programmes (calcul et tracées des spectres ; des pôles, etc) se fait avec le même programme qui est insensible à la méthode utilisée. Enfin, l'accès des programmes à plusieurs fichiers étant possible des études comparatives peuvent être entreprises avec le même passage de programme. Conçus de cette façon, ces algorithmes non seulement permettent toute anticipation, mais ils restent particulièrement ouverts à l'avenir.

8. SYNTHESE DE LA PAROLE

8.1. Généralités

Le terme "synthèse de la parole" est utilisé pour caractériser un certain nombre de techniques, souvent très différentes, de génération d'un signal vocal à partir des paramètres qui proviennent d'une analyse du même ou d'un autre signal vocal.

Le problème de synthèse serait réduit à un problème de mémorisation, et d'accès aux donnés, si on disposait de suffisament de mémoire pour enregistrer "tout ce qui peut être dit", et si, la recherche de l'information demandée, pouvait se faire , chaque fois en temps réel. Ceci n'étant pas possible, on est obligé d'envisager des représentations plus économiques. Bien sûr, dans le cas d'un dictionnaire très limité, ces restrictions ne se posent pas. Ainsi, des appareils d'enregistrement directs sont utilisés, dotés de la possibilité d'une recherche effectuée avec une procédure plus sophistiquée. Ces appareils sont souvent qualifiés de "synthétiseurs". Nous croyons, que le vocable, "unité de réponse vocale", serait plus réaliste dans ce cas. Ainsi le mot "synthétiseur", pourrait être réservé à des appareils qui utilisent des procédures plus compliquées.

Nous savons qu'une représentation numérique directe du signal vocal pourrait s'éffectuer après echantillonage (à 8000÷10000Hz) et conversion A/D de (7÷9) bits, ce qui implique un taux de transmission très élevé, de l'ordre de (56.000÷90.000) $\frac{\text{bits}}{\text{sec}}$.

Par ailleurs , si on tient compte des probabilités d'occurence des différents phonèmes dans les différents langages , ainsi que de la capacité du cerveau humain de traiter de l'information , nous pouvons conclure que la parole ne peut contenir plus de 50 $\frac{\text{bits}}{\text{sec}}$ d'information .

On arrive actuellement à pouvoir représenter le signal vocal avec un taux entre (600-9000) bits/sec suivant la méthode d'analyse adoptée . La mise en valeur de certaines"lois de phonation " pourrait donner des taux encore plus réduits .

Ainsi deux aspects sont essentiels pour la synthèse de la parole :

. Représentation économique du signal
. Synthèse à partir des unités informationnelles réduites.

A) Représentation économique du signal

Deux applications sont liées directement avec cette technique.

a) Mémorisation économique du signal dans le cadre d'une unité de réponse vocale.

b) Transmission des données

Avec les premières techniques de codage de la parole, on a essayé, tout simplement, de décrire le signal. Parmi ces techniques, on trouve la deltamodulation ainsi qu'un certain nombre de techniques plus modernes comme la "differential pulse code modulation" (DPCM) (voir /53/), etc ... Ce sont les corrélations statistiques d'échantillons adjacents de la parole qui sont exploitées. On arrive, de cette façon, à des taux d'information de l'ordre de 20 000 bits/sec. Ces méthodes présentent l'avantage de simplicité ; mais le codage n'est pas effectif.

Pour améliorer le taux d'information, il faut tenir compte de certaines propriétés de l'appareil de phonation ainsi que de l'appareil acoustique. Dans cette catégorie de techniques, on peut inclure toutes les méthodes d'analyse et de reconstitution, ainsi que toute sorte de vocoder, la synthèse à partir des formants, les méthodes d'identification du système de phonation, etc.... Parmi les différentes techniques des taux différents sont obtenus :

Les vocoders classiques à canaux peuvent donner un taux d'information de l'ordre de 2400 bits/sec (/3/).

La synthèse à partir des formants peut donner des taux très réduits de l'ordre de 600 bits/sec (/15/).

La représentation avec un modèle d'une fonction de transfert en z (coefficients de prédiction) peut donner un taux d'information de 3 000-9 000 bits/sec suivant la qualité demandée et le mode d'analyse.

(synchrone -asynchrone).

Les différences de taux d'information entre les différentes méthodes doivent être vues toujours en rapport avec la qualité de la voix synthétique obtenue et avec le volume des calculs à effectuer .

Un point intéressant de toutes ces méthodes est la possibilité d'intervention au niveau de la reconstitution pour altérer , améliorer ou modifier certaines caractéristiques . Ainsi , avec de simples homothéties on peut modifier des voix d'hommes en voix de femmes etc ..... La simplicité de modification est dûe au fait que les différents contenus informationnels se trouvent isolés après analyse . De cette façon on peut commander indépendement , les caractéristiques liées avec les formants , le pitch , la puissance et les différentes durées . Des modifications simples peuvent être envisagées . Les résultats obtenus sont assez remarquables .

Nous avons déjà vu au chapitre 3, deux vocoders basés sur des méthodes analogiques d'analyse, et deux autres basées sur la définition du cepstre.

Nous donnerons au § 8.3. les détails techniques sur un appareil qui fonctionne avec un analyseur basé sur le filtre de Kalman.

B) Synthèse à partir des unités informationnelles réduites.

Ce type de synthèse est souvent qualifiée de "synthèse par règles". On recherche en général des lois suivies à un niveau de commande supérieure, par le mécanisme cybernétique de la production de la parole. Les unités informationnelles utilisées peuvent être très différentes suivant le cas :(mots, diphonèmes, phonèmes ou autre unités élémentaires).

La simple juxtaposition des différentes unités n'étant pas suffisante pour assurer la continuité du méssage final, on recherche en général des lois de passage.
Dans le cas de la juxtaposition des mots, le problème est simple. Cette solution peut être utilisée dans le cas d'un dictionnaire limité. Elle est alors la solution la plus réaliste, vue la simplicité de l'unité de réponse vocale correspondante, et la qualité de la voix à la sortie. S'il s'agit d'un dictionnaire très étendu, la solution précédente se heurte au volume des données à mémoriser,même si ces données sont représentées d'une manière économique (coéfficients de prédiction par exemple).

Des unités plus élémentaires sont à étudier dans ce cas , comme par exemple : les diphonèmes et les phonèmes , dont le nombre est assez restreint dans le contexte d'une langue , (600 diphonèmes et 39 phonèmes en français). Mais ce que l'on gagne en mémoire , on le perd souvent en qualité de la parole synthètique . Un nombre de travaux très restreint a été présenté jusqu'à présent , dans ce domaine /8/

Les transitions phonèmiques sont à l'origine d'un grand nombre de difficultés . En effet , la juxtaposition " des états stables " n'assure pas la continuité du message parlé . (voir chapitre 9)

La synthèse par règle au niveau phonème à besoin d'un taux d'information de l'ordre de 75 bits/sec. Une telle réduction du taux d'information est nécéssaire dans le cadre d'un nombre très limité d'applications, (robot universel, encyclopédie parlante). Pourtant elle constitue un but de recherche très attirante, si la compréhension du mécanisme de phonation nous intéresse.

8.2. Modèles de synthèse.

Pour produire de la voix synthètique, il faut pouvoir commander un"modèle de production". Ce modèle peut être très différent suivant l'analyse qui a précédé. Trois types de modèles sont à distinguér.

. Modèles articulatoires (mécaniques)
. Modèles électriques
. Modèles abstraits

Les modèles articulatoires sont utilisés dans le but de se passer des problèmes crées par les transitions phonèmiques . On peut s'imaginer ainsi que le passage d'un état stable à un autre état stable pourrait se faire d'une manière naturelle . La transition donc doit se reproduire.Les mouvements d'un certain nombre d'éléments mécaniques qui simulent la langue, les lèvres etc..., sont commandées au niveau de la synthèse. Le problème qui se pose est de pouvoir lier la commande de ce système avec une analyse du signal qui a préexisté. Pour l'instant il y a très peu de travaux qui sont publiés dans le contexte d'un tel modèle. /55/. Nous voyons sur la figure 8.1 un modèle mécanique de production de sons sonores (/3/) et sur la figure 8.2 un modèle articulatoire du canal vocal.

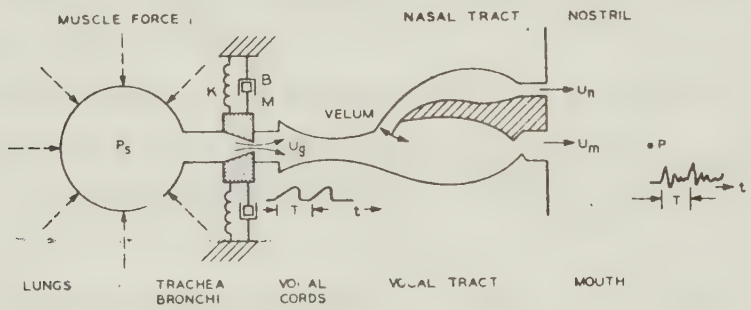


Figure 8.1 : Modèle mécanique de production de sons sonores (tiré de /3/)

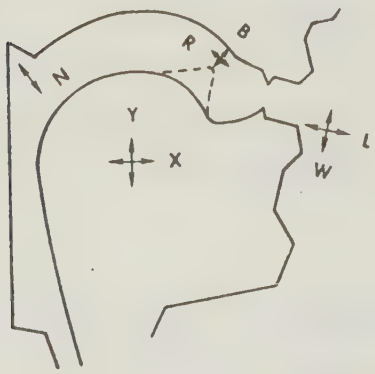


Figure 8.2 : Modèle articulatoire du canal vocal avec sept paramètres à commander (tiré de /3/).

Deux types de modèles électriques très différents existent. Le premier type est un modèle de résonateurs du second ordre mis en série ou en parallèle. On peut de cette façon approcher le caractère résonant du canal vocal. La commande de ces résonateurs peut s'éffectuer si l'on connaît les variations des formants. En ce qui concerne les amortissements, il a été démontré /56/ que leur rôle était secondaire et que la qualité de la voix synthétique était plutôt liée à l'exactitude des valeurs des fréquences.
Nous avons déjà présenté un modèle de ce genre au chapitre 3.

Un autre alternatif serait un modèle électrique qui donne l'approximation du comportement du canal vocal, considéré comme un tube acoustique: (paramètres répartis). Un tel modèle a été donné au chapitre 1.

Comme modèle abstrait on considère une méthode de génération du signal vocal, n'ayant pas de relation directe avec la nature du système de phonation. Un tel modèle sera étudié au § suivant.

8.3 Synthétiseur de la parole attaché à un analyseur KALMAN

Un modèle abstrait est utilisé pour la synthèse de la parole, sous forme d'une relation de récurrence entre les échantillons successifs du signal vocal (Voir CH 5).

$$s_n = - a_1 s_{n-1} - a_2 s_{n-2} \cdots - a_k s_{n-k} + b_o e_n \qquad (8.1)$$

Pour la commande du synthétiseur, un certain nombre de paramètres sont demandés. Ces paramètres calculés à l'analyse (filtre de KALMAN, Voir CH 7), sont les suivants :

1) le vecteur $\underline{a} = (a_1 , a_2 , \ldots , a_k)$
2) Un paramètre de voisement v (sonore - sourd)
3) La puissance du signal P, pour le calcul de b_o
4) La période fondamentale T_o

Nous voyons sur la figure (8.3), le système qui est utilisé pour la synthèse.

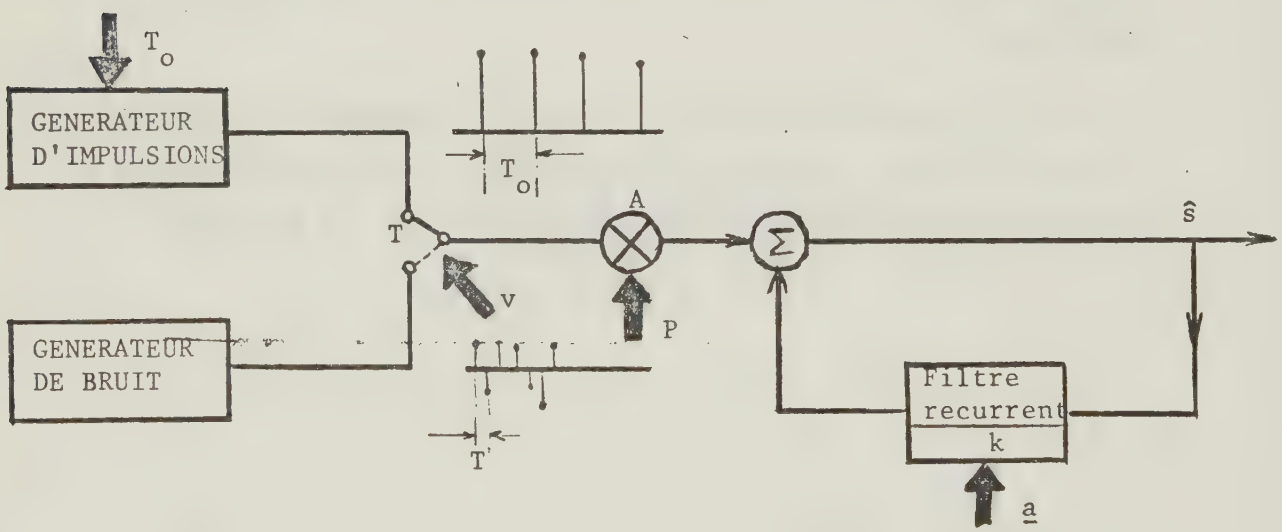


Figure 8.3 : Synthèse de la parole à l'aide d'un filtre récurrent

Les deux générateurs de la figure sont nécessaires pour exciter le système;soit un générateur d'impulsions, soit un générateur de bruit pseudo-blanc sont utilisés. Le générateur d'impulsions est commandé par la période fondamentale T_o de façon à

fournir une suite d'impulsions à une cadence égale à celle du pitch. Le générateur de bruit fournit une suite de nombres aléatoires espacés de T (T= période d'échantillonage)

le vecteur $\underline{a}$ commande le filtre k

la puissance P commande le potentiomètre A

le paramètre de voisement v commande l'interrupteur T .

Un taux d'information de l'ordre de 7000 bits/sec est envoyé au synthétiseur. Les paramètres a_i peuvent être quantifiés jusqu'à 10 bits/sec au minimum. Ces paramètres sont très sensibles aux erreurs de quantification. Ainsi, pour moins de 10 bits/sec, on peut arriver souvent à des modèles instables (pour des modèles stables au départ). Pour une petite variation $\delta\underline{a}$ du vecteur $\underline{a}$ les variations $\delta\rho$ de la distance des pôles de l'origine du plan des z sont importantes. Ainsi, on ne peut pas espérer en utilisant les paramètres a_i d'arriver à des taux très réduits (600 bits/sec par exemple dans le cas du synthétiseur à formants). En outre, le calcul soit des formants soit des pôles est trop long pour être envisagé dans le cadre de la transmission des données.

La qualité de la parole synthétique produite avec cette méthode est très bonne. Le synthétiseur est simulé sur CII-10070 en FORTRAN.

9. TRANSITIONS PHONEMIQUES .

9.1. Rôle des transitions phonemiques dans la structure d'une phrase

L'introduction du phonème comme unité linguistique de base a répondu au besoin d'établir un code pour représenter le message parlé analogue au code de représentation du message écrit , l'unité de base pour le message écrit étant le caractère .

Pourtant une grande différence existe entre les deux représentations, le message parlé étant un code analogique , la représentation par des phonèmes n'est qu'une discrétisation de ce code , et par conséquent une approximation . Cette approximation se révèle assez grossière vu le rôle que jouent les transitions phonèmiques dans le message parlé . En effet ces transitions phonèmiques , exprimant l'évolution des paramètres d'un état stable à un autre état stable , sont très informatives , car une infinité de trajectoires entre deux régions stables seraient théoriquement possibles . En outre ,un phénomène de voisinage se manifeste souvent , ce qui implique un conditionnement de l'état stable en question par les états voisins . Ainsi , chaque état stable se présente sous un éventail d'aspects différents suivant le voisinage .

Nous avons vu (chapitre I) qu'on pouvait regrouper l'information contenue dans une phrase parlée sur différents supports fondamentaux :

- variations d'amplitude du signal vocal (enveloppe énergétique)
- variation du fondamental
- régions stables d'un phonème
- régions de transition entre phonèmes.
- durée des états stables ou transitoires.

Dans le but d'une analyse du signal , les régions transitoires sont en général découpées en plusieurs segments . Chaque segment est considéré comme stable . C'est la suite de ces segments qui représente

la transition . Donc les zones transitoires sont étudiées exactement comme les zones stables . Pourtant si l'on ne peutpas étudier ces zones par des méthodes différentes (modèles non stationnaires par exemple) il y a un grand intérêt à les détecter et à les représenter par un formalisme adapté .

C'est dans ce sens qu'est dirigée cette étude qui propose l'établissement d'un modèle absrait de l'évolution transitoire de paramètres , d'origine quelconque , supposés représentatifs du signal vocal . /41/.

On notera que , malgré l'importance du problème , un nombre très retreint de travaux y a été consacré . La raison peut en être recherchée dans l'introduction relativement récente des méthodes de modèlisation dans le traitement du signal vocal . RABINER , cependant , en vue de la synthèse vocale , propose l'ajustement d'une équation differentielle du second ordre avec amortissement critique (1 paramètre libre) pour décrire l'évolution des formants lors d'une transition , mais ce modèle demeure rudimantaire .

9.2. Modèlisation des transitions.

Soit $\underline{d}_n$ un vecteur représentant les paramètres issus d'une méthode d'analyse quelconque , à l'instant n . Un modèle absrait caractérisant l'évolution temporelle de $\underline{d}_n$ est constitué par une relation de recurrence d'ordre k :

$$\underline{d}_n + \omega_1 \underline{d}_{n-1} + \ldots + \omega_k \underline{d}_{n-k} = 0 \qquad (1)$$

dont il s'agit de déterminer les coefficients ω_i au mieux . L'équation (1) ne pouvant être identiquement vérifiée sur la longueur de l'échantillon , on est amené à concéder une erreur $\underline{e}_n$ dont la matrice de covariance R_n est supposée connue .

$$\underline{d}_n + \omega_1 \underline{d}_{n-1} + \ldots + \omega_k \underline{d}_{n-k} = \underline{e}_n$$

$$E(\underline{e}_n) = 0 \quad , \quad E(\underline{e}_i \cdot \underline{e}_j^T) = 0 \quad , \quad E(\underline{e}_n \cdot \underline{e}_n^T) = R_n \tag{2}$$

Le critère d'optimalisation retenu est le minimum de variance du vecteur ω_n estimé par :

$$C = \text{Trace } P_n \quad , \quad P_n = E\left[(\underline{\omega} - \hat{\underline{\omega}}_n)(\underline{\omega} - \hat{\underline{\omega}}_n)^T \right] \tag{3}$$

La solution du problème est alors fournie par un filtre de Kalman multidimensionnel défini par l'estimateur linéaire récurrent :

$$\hat{\underline{\omega}}_n = \hat{\underline{\omega}}_{n-1} + K_n \left(\underline{d}_n - D_n^T \hat{\underline{\omega}}_{n-1} \right) \tag{4}$$

avec $$D_n^T = \left[\underline{d}_{n-1} , \underline{d}_{n-2} , \ldots , \underline{d}_{n-k} \right]$$

dont le coefficient matriciel K_n est ajusté de manière optimale par :

$$P_n = (I - K_n D_n^T) P_{n-1}$$

$$K_n = P_{n-1} D_n (R_n + D_n^T P_{n-1} D_n)^{-1} \tag{5}$$

L'initialisation des équations (5) est réalisée à un instant choisi par :

$$\hat{\omega}_0 = 0 \quad \text{et} \quad P_o = \left\{ P_{ij} : \sigma^2_P \exp\left(-\frac{1}{2} \left(\frac{i-j}{\theta} \right)^2 \right) \right\}$$

La décroissance systèmatique de la trace de P_n traduit la convergence et l'apport d'information résultant de la nouvelle mesure disporible $\underline{d}_n$.

9.3. Application à la segmentation de la parole .

Une approche analytique de la parole est nécessaire pour toute étude de quelque ambition en synthèse ou en reconnaissance vocales . Il est , en effet , indispensable quand le traitement d'un vocabulaire important est requis , de décrire et de mémoriser les caractéristiques du signal à partir d'unités de dimension suffisamment restreinte pour ne pas aboutir à un volume de données prohibitif . La détermination et l'enchainement de ces unités devient alors le problème crucial .

Ces unités d'origine acoustique auront l'avantage , bien qu'à l'évidence ceci ne soit pas indispensable /42/ ,de coincider avec des éléments linguistiques classiques pour bénéficier de l'acquis de cette discipline . La nature et la qualité de la segmentation est donc avant tout conditionnée par la pertinence des paramètres extraits du signal Un paramétrage idéal ferait apparaître sans équivoque des régions de stabilité caractérisant la présence sous- jacente d'une forme acoustique Un tel paramétrage serait inaccessible ou simplement trop coûteux à mettre en oeuvre .

Il en résulte que les méthodes usuelles de segmentation se contentent de rechercher des zones de stabilité approximatives en procédant ou non sur l'ensemble des paramètres /43/ /44/ , à partir de seuils de discrimination ajustables ou non /45/ . On a alors , à faire à un problème de détection de forme .

Une approche fine du problème est constituée par la segmentation après reconnaissance où chaque occurrence d'un vecteur de paramètres se voit appliquer une procédure de reconnaissance complète destinée à vérifier celle-ci. Une telle méthode entraine un temps de calcul très important . C'est pourquoi , on a le plus souvent recours à la détermination de diverses clefs simples , permettant le classement grossier en catégories phonétiques /46/, /47/ .Il s'agit là d'une reconnaissance élémentaire parfaitement adaptée à l'implantation d'une procédure de reconnaissance hiérarchique , mais qui se doit d'être rigoureusement exacte.

La méthode proposée ici est intérmédiaire car elle réalise une segmentation par classification différentielle et par détection des régions transitoires dans l'ensemble des paramètres . Supposant en effet la convergence obtenue , le caractère récurrent de la technique de modélisation permet de surveiller l'évolution des paramètres $\underline{d}_n$ par prédiction

$$\tilde{\underline{d}}_n = - \sum_{i=1}^{k} \omega_i \, \underline{d}_{n-i} \quad ; \quad \underline{\varepsilon}_n = \underline{d}_n - \hat{\underline{d}}_n$$

Par ailleurs , le vecteur $\hat{\omega}_n$ étant caractéristique d'une zone temporelle dont les paramètres respectent une certaine cohérence un critère intrinsèque du type :

$$\gamma_n = (\hat{\underline{\omega}}_n - \hat{\underline{\omega}}_{n-1})^T (\hat{\underline{\omega}}_n - \hat{\underline{\omega}}_{n-1}) \qquad (7)$$

rend compte du franchissement de la frontière d'une classe sans qu'il soit utile d'identifier celle-ci .

Cette procédure a été mise en oeuvre avec succès sur différents types de paramètres : canaux de vocoder , formants calculés par FFT , paramètres issus de la modèlisation du canal vocal (voir /33/). La figure 1 donne un exemple d'application dans ce dernier cas . L'ordre k du système est limité à 3 et un nouveau jeu de paramètres $\underline{d}_n$ est fourni

en synchronisme avec les périodes fondamentales pour les sons non-voisés De façon à montrer le pouvoir discriminant de l'algorithme , on a omis de le réinitialiser après chaque détection de transition . La désensibilisation du critère de segmentation , dû à l'intégration par le filtre d'événements de natures différentes ne survient qu'après quelques secondes de parole . Il apparaît cependant que l'exactitude du modèle sur les transitions rapides est conditionnée par le nombre d'échantillons disponibles . On peut considérer qu'une convergence acceptable est obtenue pour un nombre d'échantillons supérieur à 3.k.

Il est à noter , d'autre part , qu'une telle procédure est succeptible de compléter les informations issues d'une analyse rudimentaire du signal . Les paramètres du modèle traduisent une loi d'évolution des composantes de $\underline{d}_n$, dont l'existence ne saurait être soupçonnée par le simple examen des courbes temporelles . Les coefficients ω_i constituent ainsi , des paramètres d'ordre supérieur qui seraient adjoints avec profit aux paramètres bruts précédents .

Figure 9.1 : Détection des transitions SANS réinitialisation de l'algorithme. Paramètres calculés suivant la méthode du CH 7.

9.4. Conclusions

La modèlisation des transitions phonémiques par un filtre de Kalman multidimensionnel a donné lieu à une technique de segmentation qui se caractérise par :

- une application en ligne sur un jeu de paramètres quelconque.
- une prise en compte globale de l'évolution de tous les paramètres.
- un critère de segmentation lié au modèle et opérant donc une classification

Mais l'étude déborde largement ce cadre grâce à des modifications mineures . En effet l'utilisation du lissage conjointement avec la prédiction (ici seule utilisée) engendrerait un modèle tenant compte à la fois du passé et du futur de la transition . Un tel modèle serait bien adapté à la génération automatique des transitions dans le cadre de la synthèse par règles au niveau phonème.

Bien plus , la méthode d'identification adoptée , constitue une approche directe au problème de reconnaissance . L'application de l'algorithme en période d'apprentissage , fournit un modèle de repartition des occurrences sous la forme des caractéristiques d'un bruit multidimensionnel coloré . Il faut noter ici l'aspect " séquentiel " d'un tel algorithme employé dans le cadre de la reconnaissance des formes . Cet aspect se traduit par la capacité de la matrice P de covariance , de " mémoriser un tel évènement d'une manière naturelle . Ainsi au niveau décision l'apprentissage peut continuer , dotant ainsi l'algorithme de la possibilité de suivre les variations éventuelles des classes .

C O N C L U S I O N

Au cours de cet exposé un certain nombre de techniques d'analyse numérique du signal vocal ont été étudiées. Il s'agit surtout de techniques basées sur des procédures assez sophistiquées pour la recherche de l'information.

Nous avons essayé de considérer l'analyse de la parole à travers ses différentes applications. Ceci parce que sa qualité conditionne tout système de reconnaissance, de synthèse vocale, d'identification des locuteurs ou de transmission des données.

Parmi les différentes applications, nous avons étudié plus en détail la synthèse vocale. Elle a été souvent vue comme une application du type "vérification de la qualité de l'analyse". Les principaux aspects de la synthèse ont été discutés et nous avons noté les divers modèles utilisés. Notre modèle de synthèse a été donné. Son caractère abstrait sous forme d'une relation de récurrence entre les échantillons du signal vocal a été évoqué.

Nous avons exposé ici notre façon de concevoir les différentes méthodes d'analyse qui existent. Une méthode originale est développée. Dans son contexte, un modèle du système de phonation est ajusté pour fournir à sa sortie un signal qui ressemble le plus possible au signal vocal réel au sens d'un critère d'optimalité. Ce critère est le minimum de variance des paramètres du système calculé. L'ajustement du modèle se fait en utilisant les résultats de la théorie du filtre de KALMAN discret. Cette méthode est très différente de celles utilisées jusqu'à présent. Ses trois aspects intéressants sont :

1) Le caractère récurrent : Ainsi une estimation du système peut être fournie à chaque instant, ce qui donne à l'algorithme la possibilité de détection instantanée des "évènements" quand l'état estimé est assez loin de l'état réel. De cette façon, des signaux riches en évènements comme celui de la parole, peuvent être suivis régulièrement.

2) Les nombreuses solutions sous-optimales qu'elle peut fournir, le choix de l'une d'elles correspondant au choix du jeu des paramètres d'initialisation.

3) La possibilité d'obtention d'un aperçu minutieux du signal. Le modèle calculé ici reflète des caractéristiques spectrales à très court terme, qu'on ne peut pas obtenir avec d'autres techniques.

L'inconvénient de la méthode reste le grand nombre de calculs à effectuer. Ainsi, le temps réel peut être atteint uniquement avec un appareil en hardware d'une structure dont la particularité consiste en l'utilisation de plusieurs unités de multiplication travaillant en parallèle. Pour remédier au volume important de calculs à effectuer, nous avons pensé à l'utilisation des méthodes "stochastiques" pour l'estimation du système (caractère récurrent conservé). Un certain nombre des méthodes "rapides" ont été étudiées. Elles se sont avérées très sensibles aux caractéristiques des bruits impliqués. Bien que des temps de convergence assez longs aient été constatés, il semble que cette étude doive se continuer dans le sens de la détermination des bonnes propriétés de convergence.

Nous avons discuté l'importance du taux d'information, particulier à chaque méthode, au niveau philosophique et au niveau pratique. Nous avons vu que les techniques étudiées ici arrivaient à des taux très réduits par rapport à la représentation directe du signal ou sa simple description. Ceci est dû à la mise en valeur de certaines propriétés de l'appareil de phonation ou de l'appareil auditif. Pourtant, il serait souhaitable de pouvoir arriver à des taux encore plus réduits. Il semble qu'on puisse y parvenir en mettant en valeur certaines "lois de phonation" (lois liées avec le mécanisme cybernétique de production de la parole).

C'est dans cette optique que nous avons essayé de mettre en évidence de telles lois d'une façon abstraite en utilisant un modèle de transitions phonémiques. Ceci est fait a posteriori (après analyse) en utilisant l'évolution des paramètres qui représentent le signal. Des paramètres d'origine très différente ont été testés. L'application immédiate de ce modèle a donné une méthode de segmentation automatique de la parole.

Il serait très intéressant de pouvoir introduire l'aspect "segmentation" au niveau analyse du signal. On pourrait être amené ainsi à traiter différemment les régions stables et les transitions. Le prélèvement d'informations par le signal vocal se fait actuellement avec une fréquence qui n'est pas forcément la bonne. Les segments d'analyse égaux à T_o (période fondamentale) ou à 10 ms, sont en effet beaucoup trop courts pour les régions stables et beaucoup trop longs pour les régions transitoires. Un prélèvement avec un pas variable serait éventuellement désirable. C'est dans cette voie que nos études se dirigent actuellement.

Un grand progrès a été réalisé les dix dernières années dans le domaine de la parole. Ce progrès couvre aussi bien la compréhension du mécanisme de phonation que ses diverses applications. La recherche est basée sur des techniques comme l'analyse en fréquences, l'identification des systèmes, la classification et la reconnaissance des formes et leur adaptation aux problèmes particuliers. L'étape analyse du signal demeure toujours une étape où tout effort d'adaptation ou d'invention est bien accueilli.

BIBLIOGRAPHIE

/1/. G.FANT : " Acoutic theory of Speech production ", Mouton - the Hague . PARIS 1970.

/2/. G. PETERSON : "Automatic Speech Recognition procedures " , Language and Speech , pp 200 - 219 , 1961

/3/. J. FLANAGAN : "Speech Analysis , Synthesis and Perception ". second Edition . Springer - Verlag New York , 1972

/4/. A.M. NOLL : " Short - time spectrum and "Cepstrum " technique for vocal pitch detection." Journal of the Acoutical Society of America . 36, pp 296 - 302 .

/5/. F. ITAKURA , S.SAITO :"An analysis - synthesis telephony based on maximum likelihood method." Proc. Int . Congr . Acoust. C-5-5 Tokyo - 1968

/6/. B. GOLD : " Computer program for pitch extraction " Journal of the acoustical Society of America . 34, pp 916 -921, 1962

/7/. A. MAISSIS :" Méthode d'extraction du pitch " L'Onde Electrique , Février 73

/8/. J.S. LIENARD : " Analyse , Synthèse et reconnaissance automatique de la parole " . Thèse de Doctorat d'Etat , Univ. Paris VI (21 Avril 1972).

/9/. R.C. LUMMIS : " Speaker Verification : Step toward the " Checkless " Society " , Bell Labs Record . Sep. 1972

/10/. B.S. ATAL, S.L HANAUER : " Speech Analysis and Synthesis by linear prediction of the speech wave " .
Journal of Acoutical Society of America , 50 , pp 637-655 , 1971.

/11/. Fast Fourier Transform : " Special issue "
IEEE trans on AUDIO and ELECTROACOUTICS , Vol AU-15 , June 1967

/12/. F. FIEVET , K. NGUYEN , V. ROUSSEL : " Programme numérique d'estimation des densitées spectrales "
Convention D.G.R.S.T. 6801242 , Tome III , Dec 1969 .

/13/. A.M. NOLL : " Cepstrum pitch determination "
Journal of Acoustical Society of America , pp 293-309, 1966.

/14/. J.D. MARKEL : " The SIFT Algorithm for fundamental fréquency estimation " . IEEE Trans . on audio avec Electroacoutics, vol. AU-20 , NO.5. , December 1972.

/15/. R.W. SHAFER , L.R. RABINER . " System for Automatique formant Analysis of voiced Speech ".
Journal of Acoutical Society of America . 47 , Part II , pp 634-648 1970.

/16/. L.R. RABINER , R.W. SCHAFER, C.M RADER : " Chirp Z - Transform algorithm and its applications "
The Bell Syst. Technical Journal , pp 1249-1291 , May - June 1969.

/17/. J.P OLIVE : " Automatic Formant tracking by a Newton- Raphson technique " . The Journal of the Acoust. Soc. of America , Vol 50 NO.2. (Part 2) 1971.

/18/. A. PAPOULIS : " Probability , Random Variables and Stochastic Processes" M.C GRAW -HILL .

/19/. M.R. ITO , R.W. DONALDSON : " Zero crossing measurements for analysis and recognition of Speech sounds " .
IEEE Trans . Audio and Electroac, Vol AU-19, 3,
p 235 -242 , 1971.

/20/. A. MAISSIS , L.F. PAU , C.J GUEGUEN : " Parametrisation et procedures de reconnaissance de la parole " . Automatisme - Tome XVII , N° 3 Mars 1972.

/21/. R. DE . MORI : " Speech analysis and recognition by computer Using zero - crossing Information ". Acustica , Vol 25 , pp 269 - 279 , 1971.

/22/. A. MAISSIS , F. FIEVET , P. WALRAVE : " A new Analog digital filtering method for real time speech recognition " . IEEE Trans . on Audio and Electroacoustics . Dec 1970.

/23/. A. MAISSIS ; " Normalisation non-linéaire des paramètres phonemiques Application à la segmentation automatique ". 4ème Journée du Groupe " Communication parlée " du GALF Bruxelles 1973.

/24/. L. LEBART,J.P. FENELON: "Statistique et informatique appliquées". Edition DUNOD.

/25/. B.S. ATAL , M.R. SCHROEDER : "Adaptative coding of speech signals. B.S.T.J. Vol 49 N°8 pp 1973 -1986 - 1970.

/26/. F.REZA : " Linear Spaces in Engineering " Blaisdell Publishing Co (A division of Ginn and Co)

/27/. F. ITAKURA , S.SAITO: A statistical method for estimation of speech spectral density and formant frequencies. Electronics and communications in Japan Vol 53 A N° 1 1970

/28/. F. ITAKURA , S. SAITO : " Digital filtering techniques for speech analysis and Synthesis " 7th International Congress on Acoustic , Budapest , 1971 (25CI)

/29/. J.D MARKEL A.H. GRAY : "On autocorrelation Equation as applied to speech Analysis ".
IEEE Transactions on Audio and Electroacoutics
Vol AU - 21 , No 2 April 1973.

/30/. A. BARRAUD : Contribution à l'identification des séquences de pondération des systèmes dynamiques
Thèse de Docteur Ingénieur , Université de Nantes , 1971.

/31/. G. CARAYANNIS ; Identification automatique des locuterus .
Rapport D.E.A. (INSTN - Saclay , - ENST)
Juin 1971.

/32/. ATAL B.S. , HANAUER S.L : " Low -bit - rate speech transmission by linear prediction of speech signals " . J. Acoust.Soc Amer . 49 , 133 (A) (1971)

/33/. C. GUEGUEN , G. CARAYANNIS : "Analyse de la parole par filtrage optimal de Kalman" .
Automatisme . Tome 18 N° 3 . 1973

/34/. C. GUEGUEN , G. CARAYANNIS : "Analyse de la parole par filtrage optimal".
4ème Journée du groupe " Communication parlée du GALF
Bruxelles 1973.

/35/. C.GUEGUEN , G. CARAYANNIS : " Analyse de la parole par modelisation Recursive " soumis aux assises de " 1974 . International Zurich seminar on digital communications "

/36/. C. GUEGUEN , A. MAISSIS , LF PAU , G . CARAYANNIS : Communication homme-machine sur support vocal".
Rapport final et annexes techniques contrat CRI 71-14 . 1972

/37/ MAKSYM J.N . "Real time pitch period extraction by adaptative prediction of speech wave form "
1972 . Conference on Speech Communication Processing , Boston , April 1972.

/38/. TSYPKIN Ya.Z. " Adaptation and learning in Automatic Systems " . Academic press , Nex York - London .

/39/. " Pattern Recognition and Machine Learning "
(Proceedings of the Japan - U.S . Seminar on the Learning process in Control Systems , Held in Nagoya , Japan . August 18-20 , 1970) . Edited by K.S. Fu .
PLENUM PRESS . N.Y. - London 1971.

/40/. J. RICHALET , A. RAULT , R. POULIQUEN .
" Identification des processus par la méthode du modèle " .
Gordon and Breach , Paris - Londres - N.York.

/41/. G. CARAYANNIS : "Modèlisation des transitions phonèmiques ; Application à la segmentation automatique."
4ème Journée du groupe " Communication parlée " du GALF .
Bruxelles 1973.

/42/. W.A. LEA : "An approach to syntactic recognition without phonemics."
IEEE Trans. on Audio Vol 21 . To be published April 73

/43/. D.R. REDDY : "Segmentation of Speech sounds"
J. Acoust. Soc . Am . Vol 40 , N°2 ; pp 307 - 312 - 1966

/44/. D.R. REDDY , P.J. VICENS : "a procedure for segmentation of connected speech."
J. Audio . Eng . Soc . Oct 1968.

/45/. S.K. DAS , D.F. STANAT : "Segmentation of utterances of a known phrase using linear threshold techniques"
IEEE Trans. on Audio Vol 20 N°2 pp 142-150 , 1972.

/46/. K.P. LI , G.W. HUGHES , T.B. SNOW : "Segment classification in continuous speech"
IEEE Trans . on Audio Vol 20 N°2 pp 142-150 1972.

/47/. ALINAT : "Reconnaissance de phonèmes en temps réel en vue de réaliser des liaisons à faible débit . Application à la Stenotypie automatique".
Rapport final d'Etudes DRME N° 198 / 71 - 1972

/48/. R. KALMAN : "A new approach to linear filtering and prediction Problems."
ASME . Trans . , Journal of Basic Eng . 35 , 1960.

/49/. F.R. GANTMACHER : " Théorie des Matrices " Tome 1
Editions DUNOD , Paris 1966.

/50/. J.D. MARKEL : " Digital inverse filtering - A new tool formant trajectory estimation".
IEEE Trans on audio vol 20 N°2 pp 129 - 137 - 1972.

/51/. L.G. KERSTA : " Voiceprint Identification " , Nature 196, 1253-57 , 1962.

/52/ DAS S.K - MOHN . W.S. " A scheme for Speech Processing in Automatic Speaker Verification
(IEEE Trans . on Audio and Electroacoustics . Vol AU-19 N°1 March 1971).

/53/. E.N. PROTONOTARIOS : "Slope Overload Noise in Differential Pulse Code Modulation Systems " , Bell Syst. Tech. J., Vol 46 , pp 2119-2161 . November 1967.

/54/. RADIX "Introduction au Filtrage numérique" (Eyrolles)

/55/. FLANAGAN J. L. "Focal Points in Speech Communication Research " IEEE , Transactions on Communication Technology . Vol COM-19, N° 6, December 1971 .

/56/. J.L. FLANAGAN , C.H. COKER , L.R. RABINER , R.W. SCHAFER , N. UMEDA . "Synthetic voices for computers " IEEE-Spectrum , October 1970 .

/57/. SARIDIS G.N. - STEIN G. "Stochastic Approximation Algorithms for Linear Discrete Time System Identification " IEEE Trans on Automatic Control, Vol AC 13 N° 5 OCT. 1968.

/58/. HOGE H. "Stochastic algorithms in system identification" INT. J. CONTROL , 1973, vol. 17 N° 6 , 1121-1128 .

/59/. C. GUEGUEN , A. MAISSIS , L. F. PAU . "Communication homme-machine sur support vocal " L'écho des Recherches, October 1972.

/60/ OPPENHEIM A. V. : "Speech analysis-synthesis system based on homomorphic filtering"
J. Acoust. Soc. Amer. 45, 459-462 (1969)

/61/ OPPENHEIM A. V., SCHAFER R. W., STOCKHAM T. G. : "Non linear filtering of multiplied and convolved signals Proc."
IEEE 56, 1261-1291 (1968)

/62/ ATAL B. S.:"Automatic speaker recognition based on pitch contours (Ph. D. Thesis).

A N N E X E I

ANALYSE DE DIFFERENTS PHONEMES AVEC LA METHODE DE DECONVOLUTION DU SIGNAL VOCAL

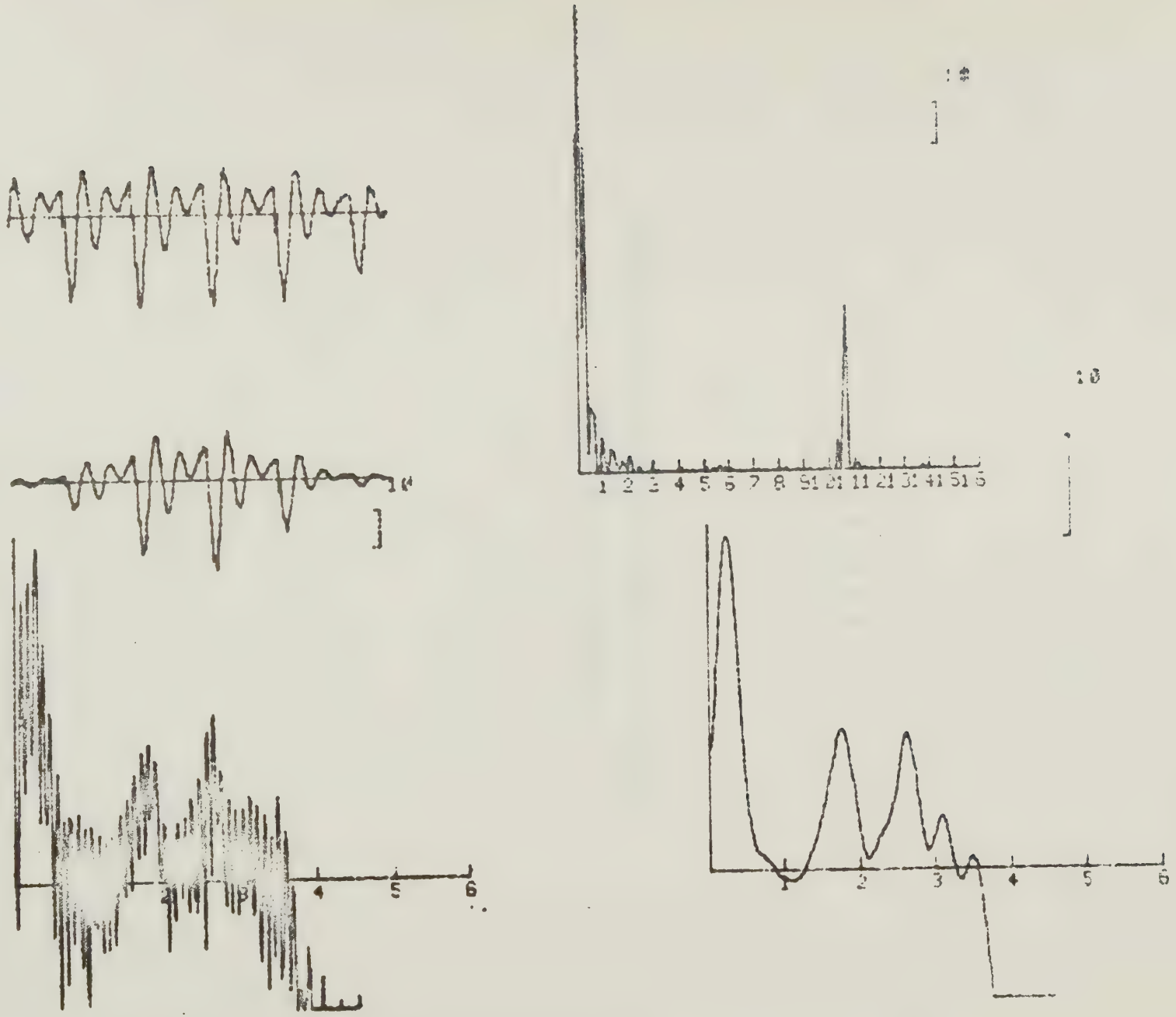


Figure A1 : Phonème /i/ .

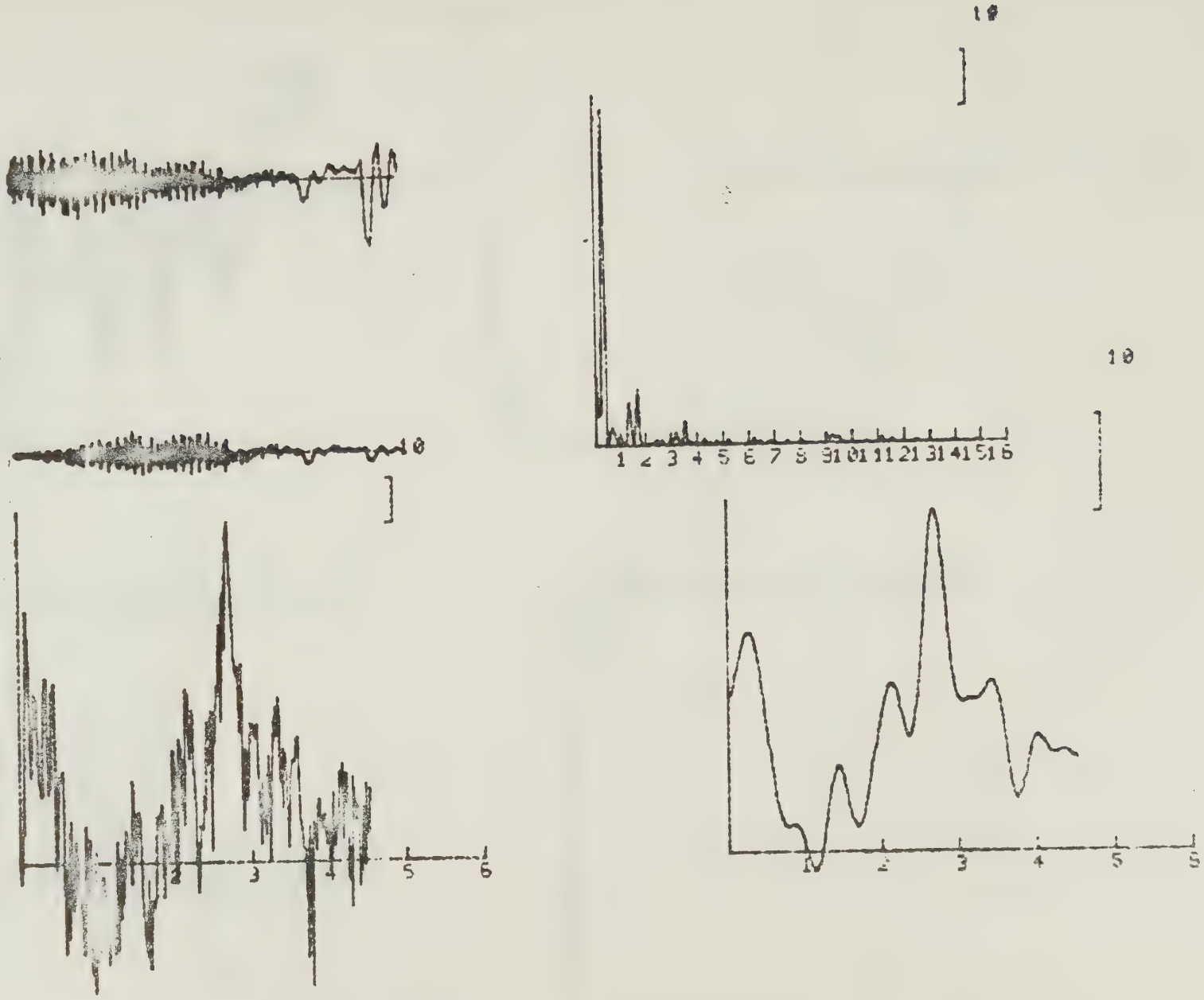


Figure A2 : Phonème /g/ .

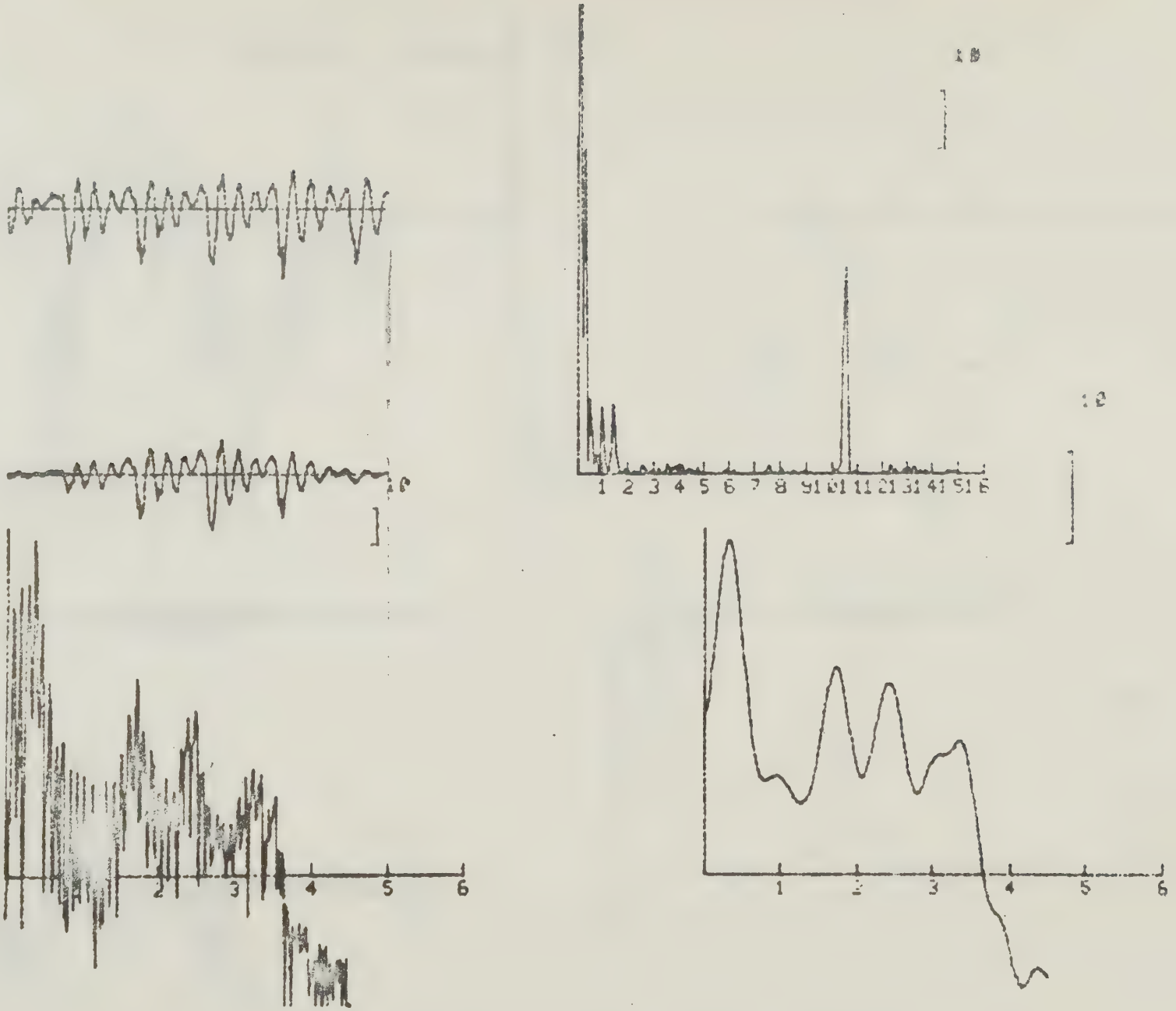


Figure A3 : Phonème /ê/ .

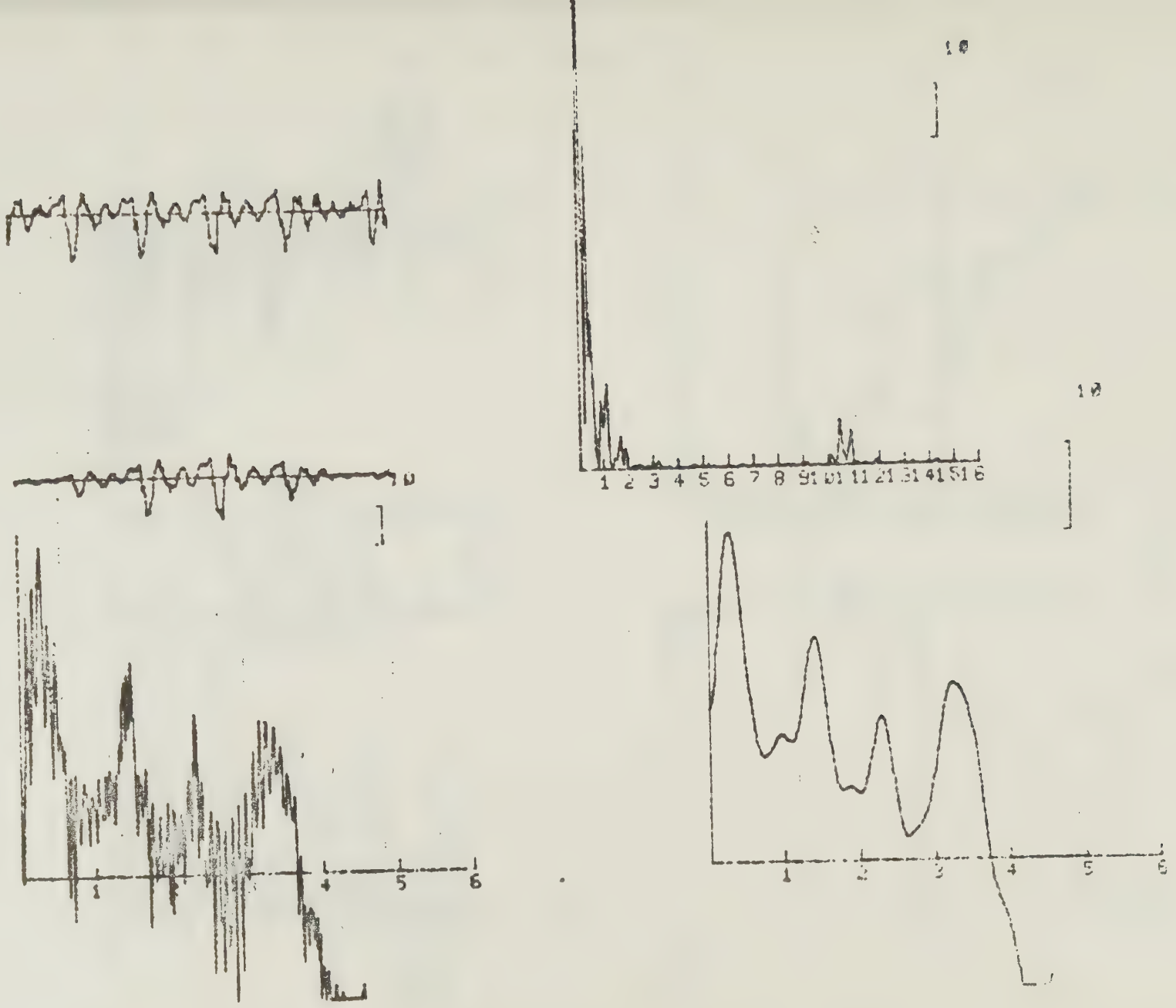


Figure A4 : Phonème /e/ .

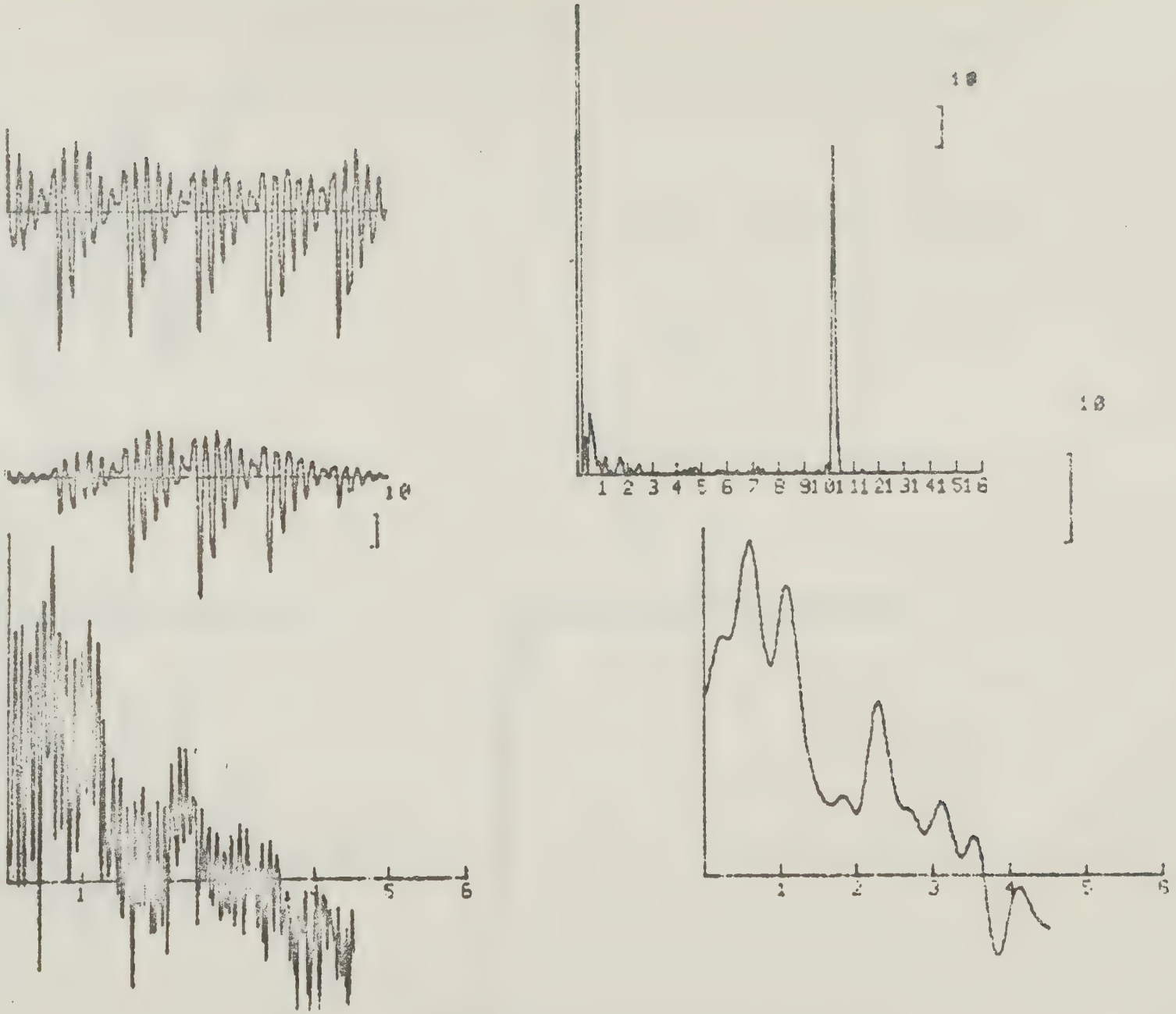


Figure A5 : Phonème /o/ ,

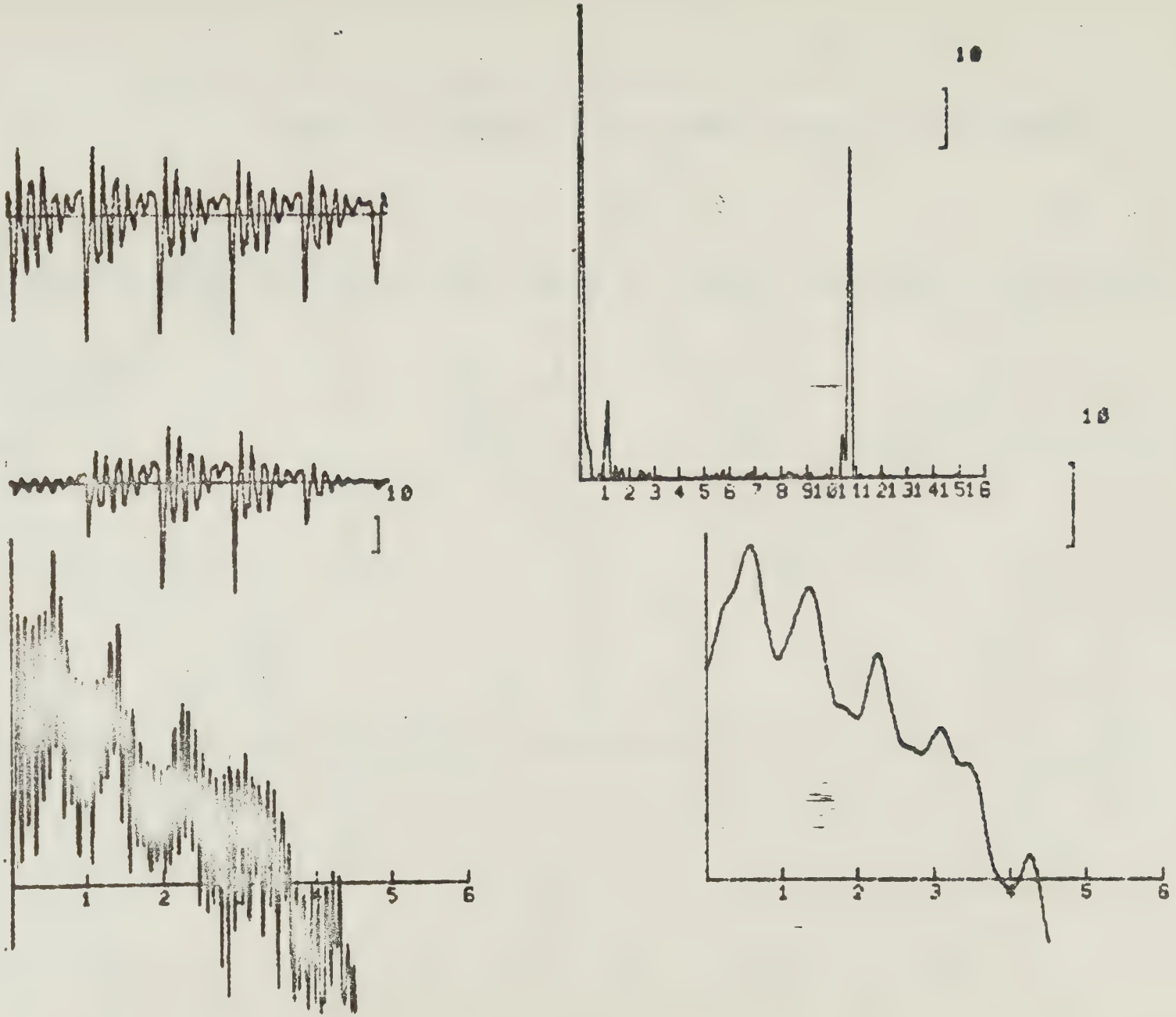


Figure A6 : Phonème /a/ .

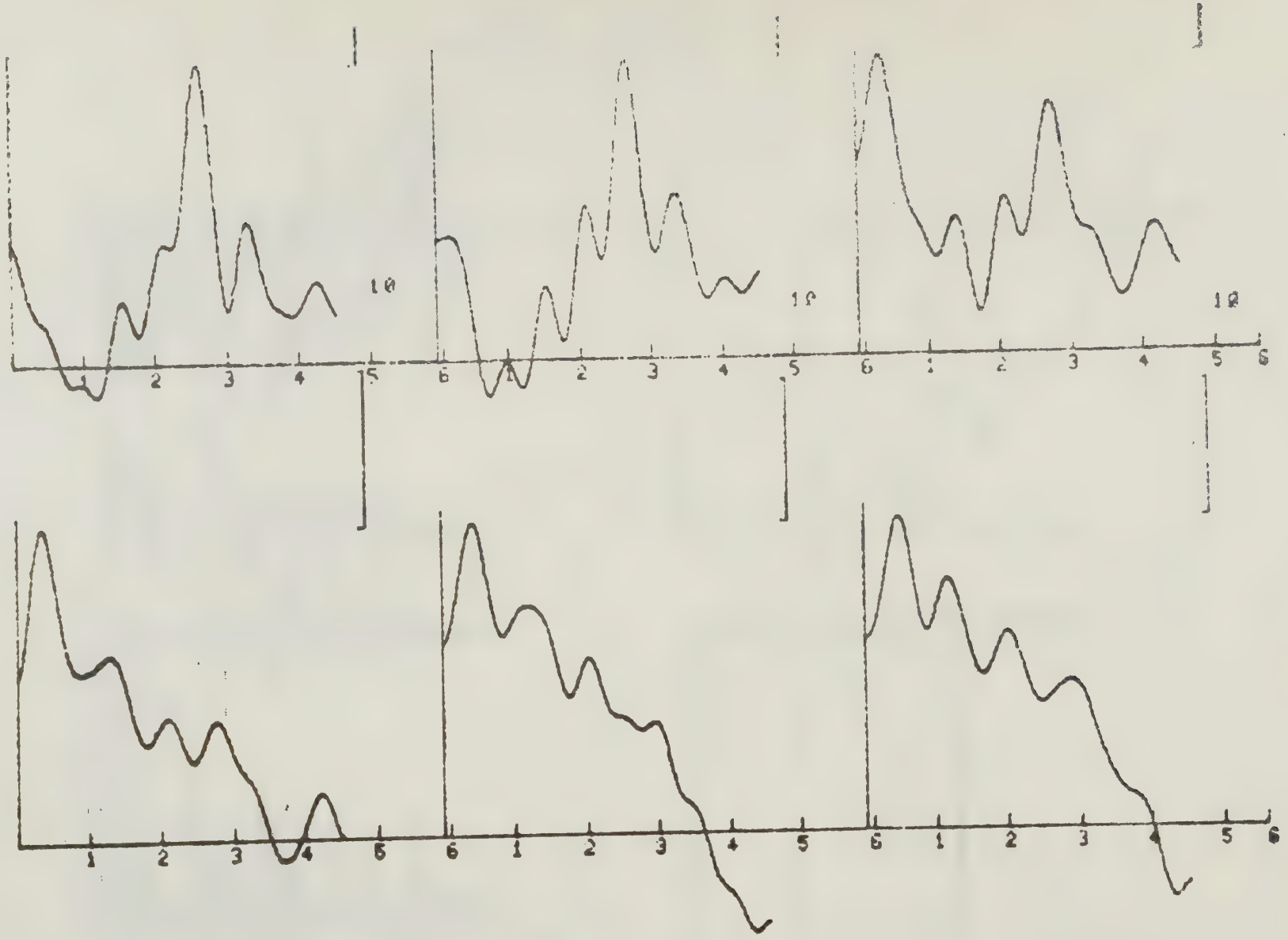
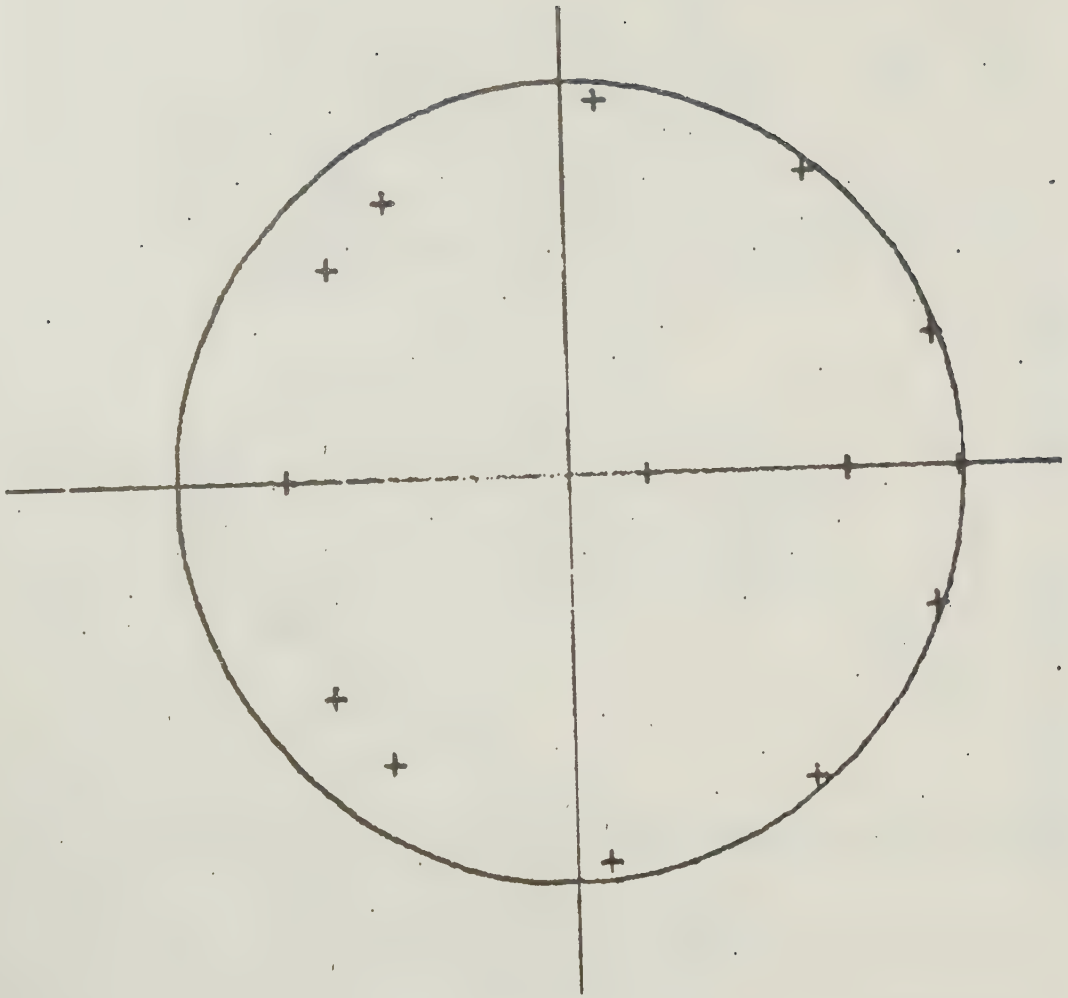


Figure A7 : Evolution de l'enveloppe du spectre sur la transition :
(/g/ → /an/) .

A N N E X E II

ANALYSE DE LA PAROLE AVEC LE FILTRE DE KALMAN



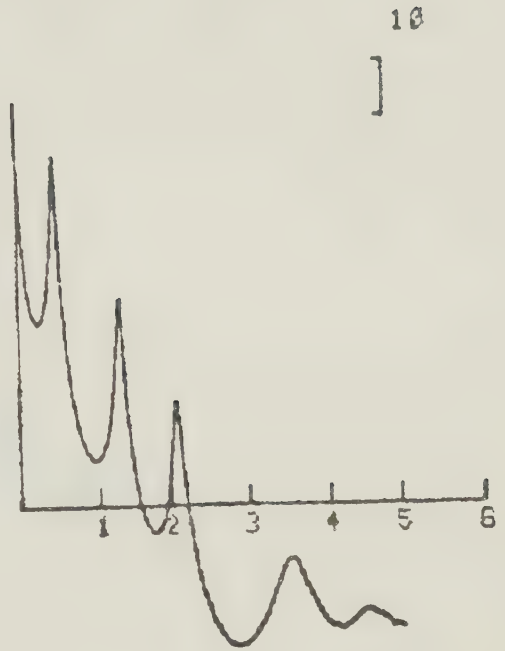


Figure A8: Pôles et spectre du phonème /a/ calculés avec le filtre de Kalman.

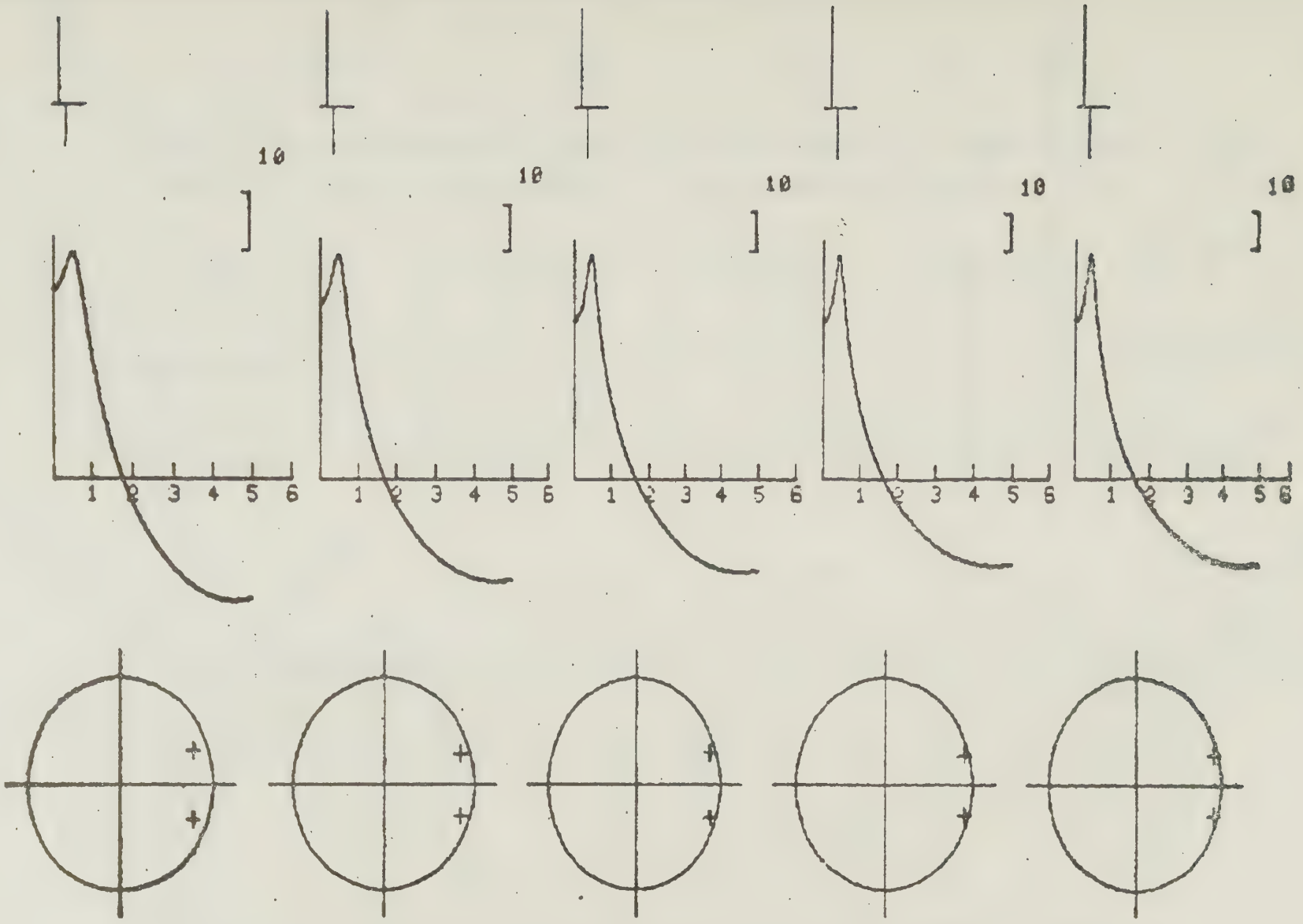


Figure A9: Evolution des parametres a_i du spectre et des pôles sur une region stable d'un phonème pour k=2 (ordre du système)

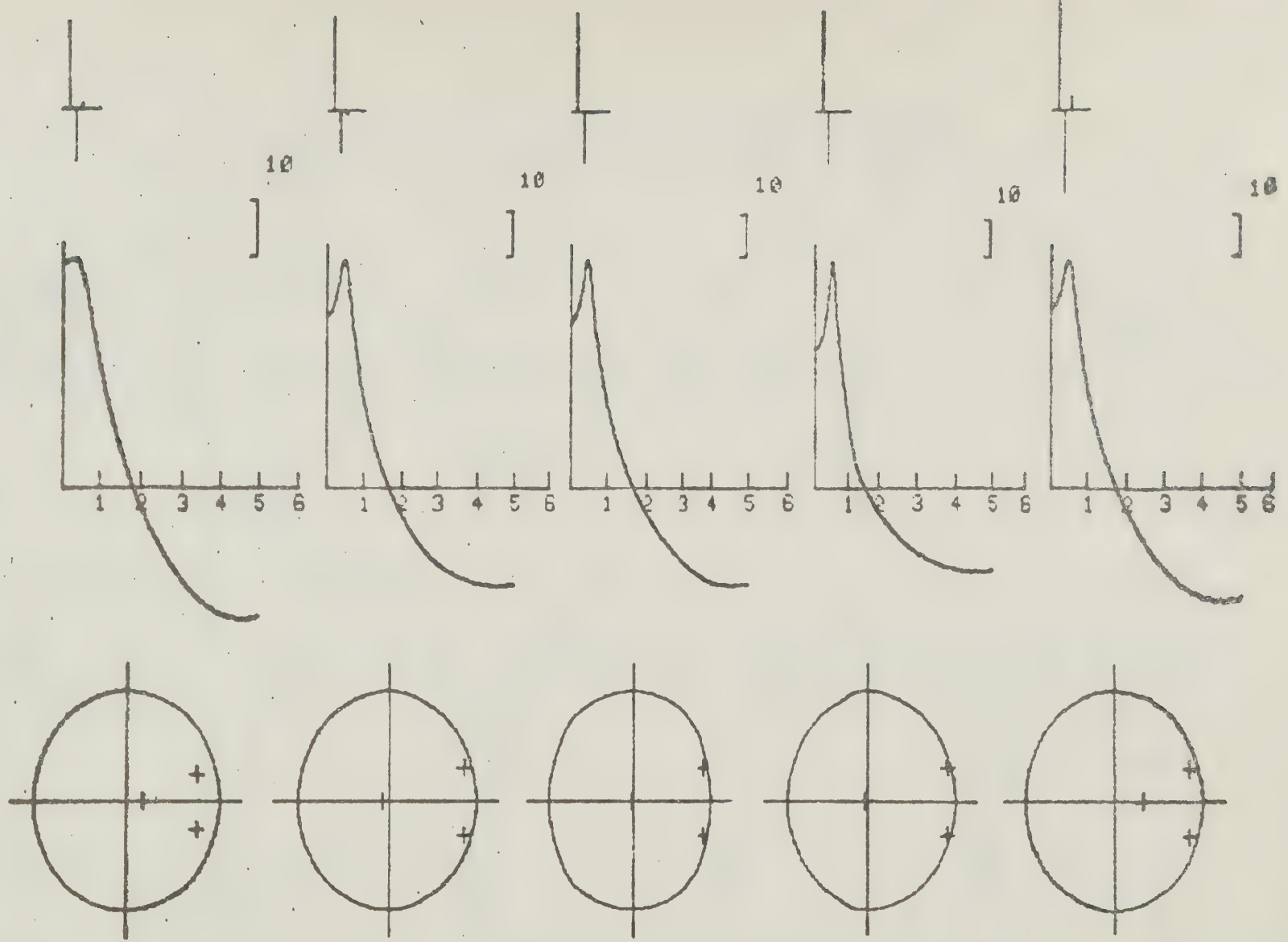


Figure A10 : Evolution des paramètres a_i, du spectre et des pôles sur une region stable d'un phonème pour k=3. (calcul effectué avec le filtre de Kalman).

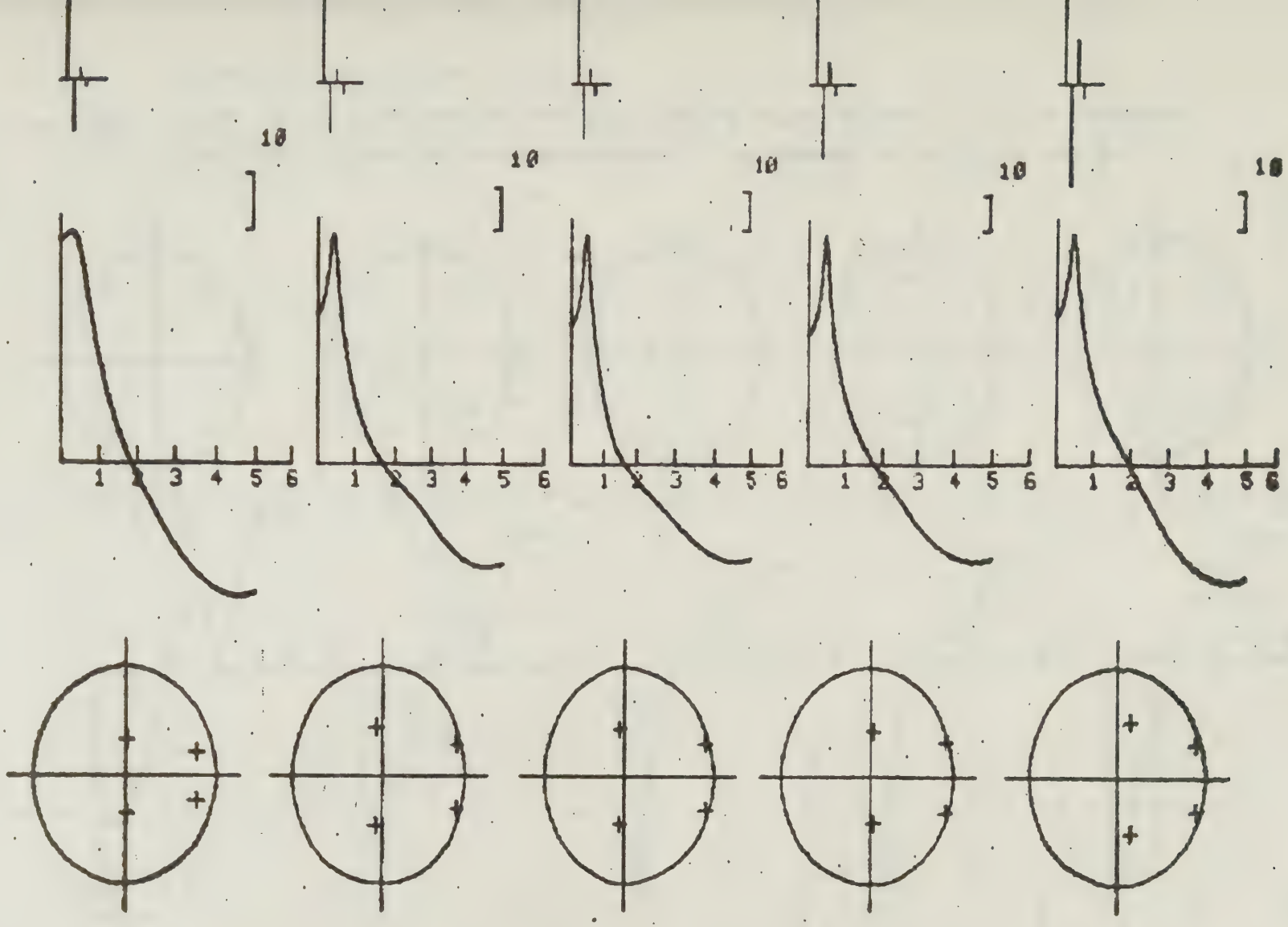


Figure A11 : Evolution des paramètres a_i , du spectre et des pôles sur une région stable d'un phonème pour k=4. (calcul effectué avec le filtre de Kalman). période d'echantillonage T=110µsec .

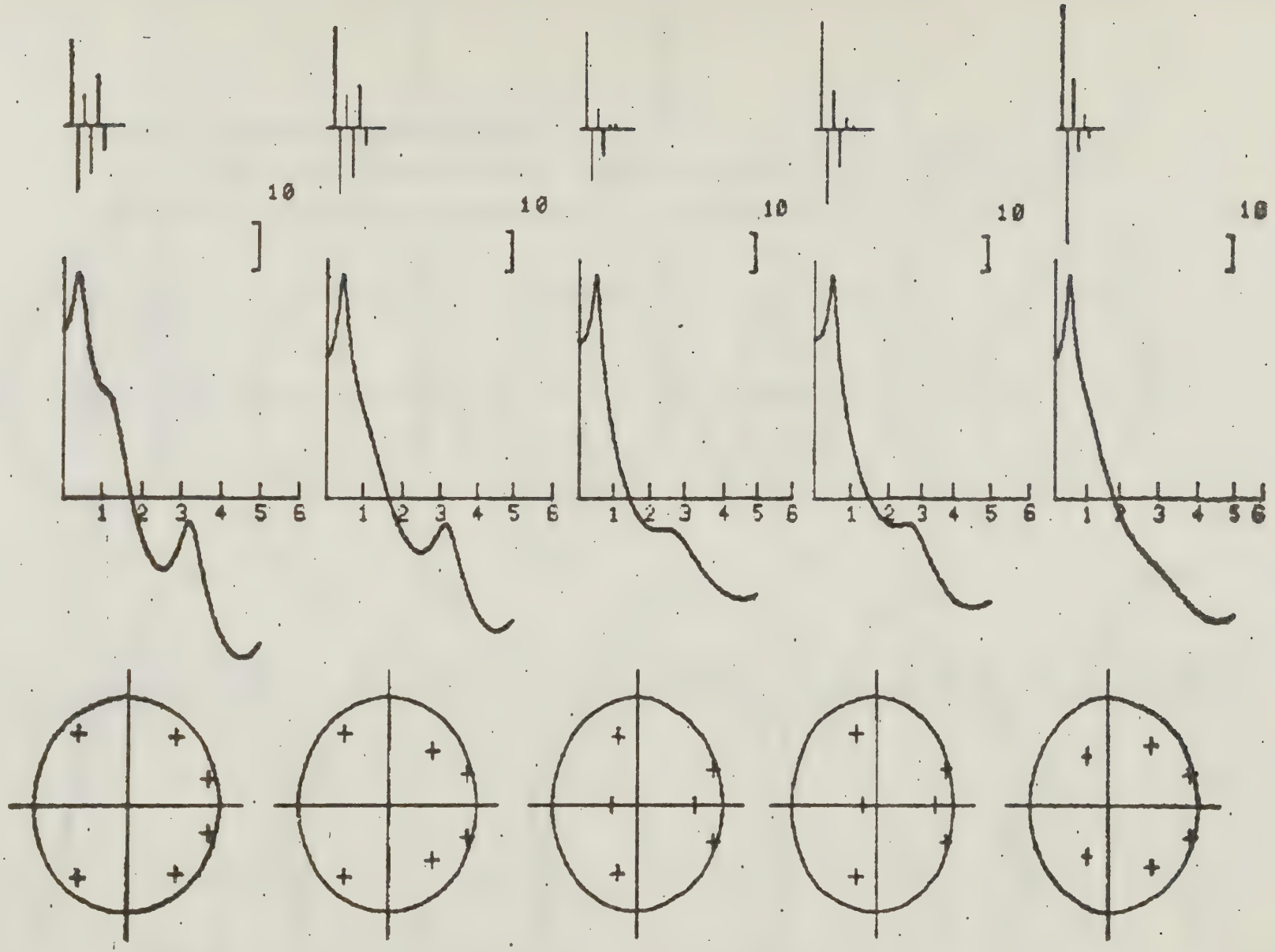


Figure A12 : Evolution des paramètres a_i , du spectre et des pôles sur une région stable d'un phonème pour k=6 .(calcul effectué avec le filtre de Kalman) période d'echantillonage T=110µsec .

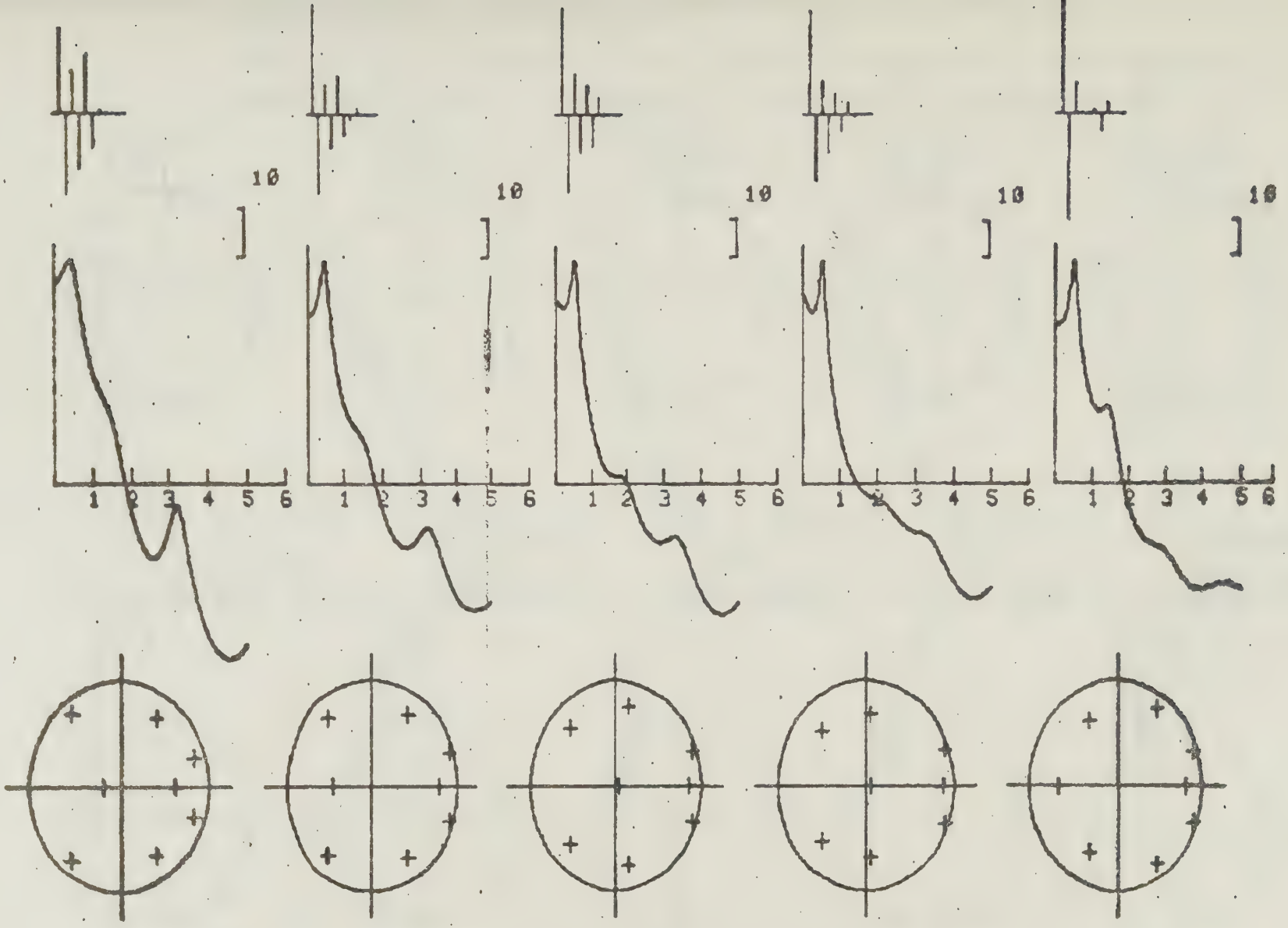


Figure A13 : Evolution des paramètres a_i , du spectre et des pôles sur une région stable d'un phonème pour k=8 . (calcul effectué avec le filtre de Kalman). période d'echantillonage T=110µsec .

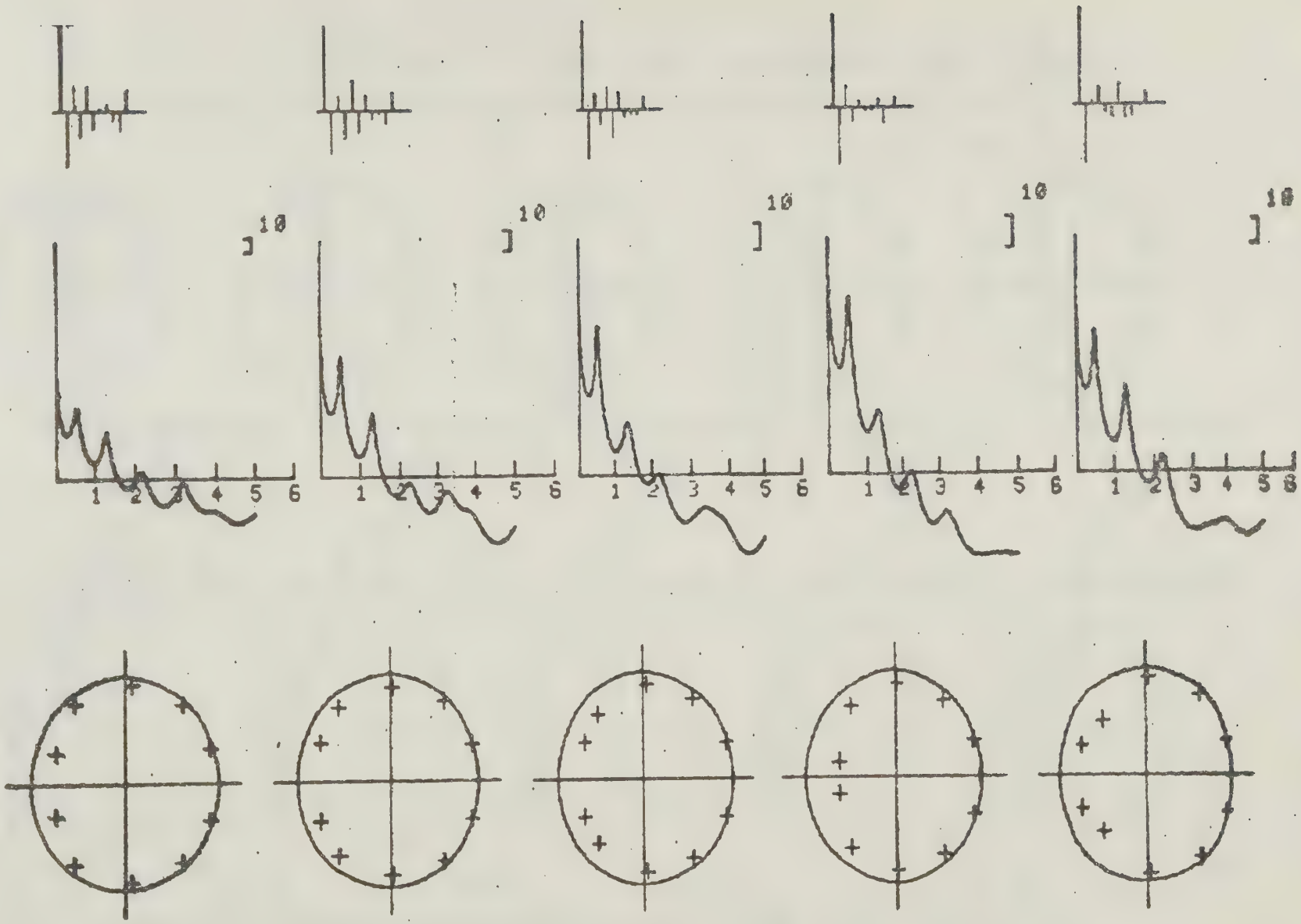


Figure A14 : Evolution des paramètres a_i , du spectre et des pôles sur une région stable d'un phonème pour k=11 . (calcul effectué avec le filtre de Kalman). période d'echantillonage T=110µsec .

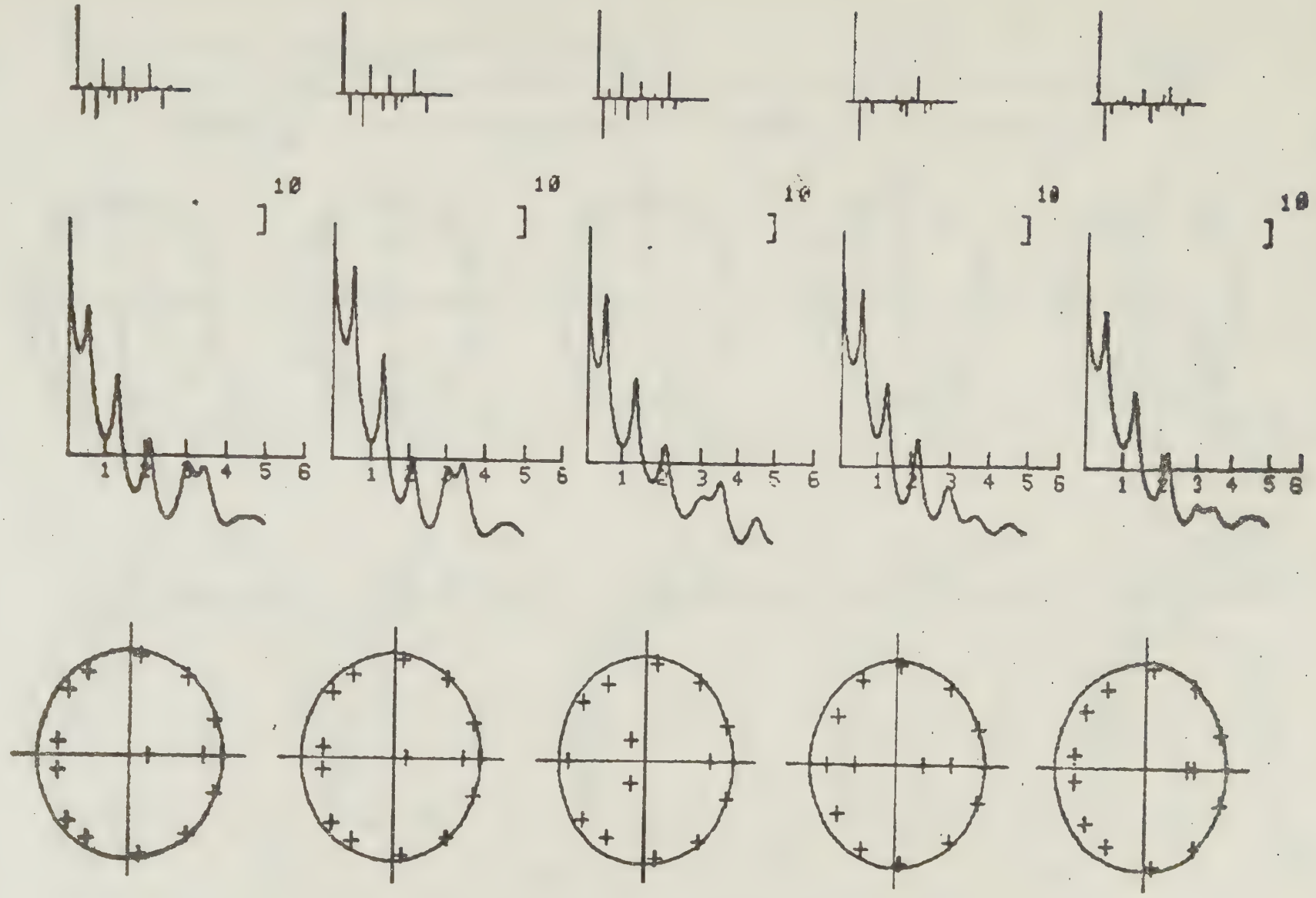


Figure A15 : Evolution des paramètres a_i ,du spectre et des pôles sur une région stable d'un phonème pour k=15 . (calcul effectué avec le filtre de Kalman) période d'echantillonage T=110µsec .

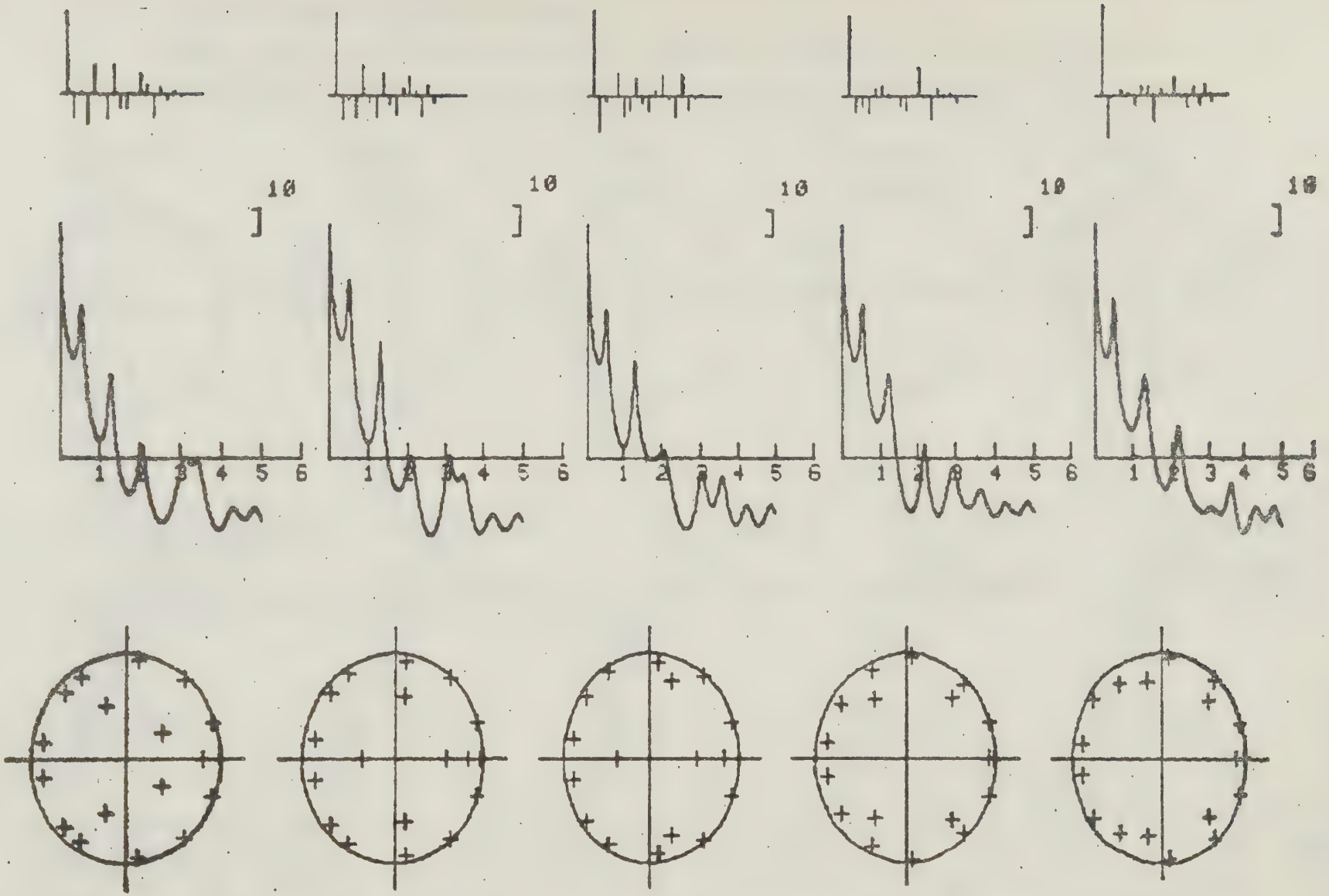


Figure A16 : Evolution des paramètres a_i , du spectre et des pôles sur une région stable d'un phonème pour k=18 . (calcul effectué avec le filtre de Kalman). période d'echantillonage T=110µsec .

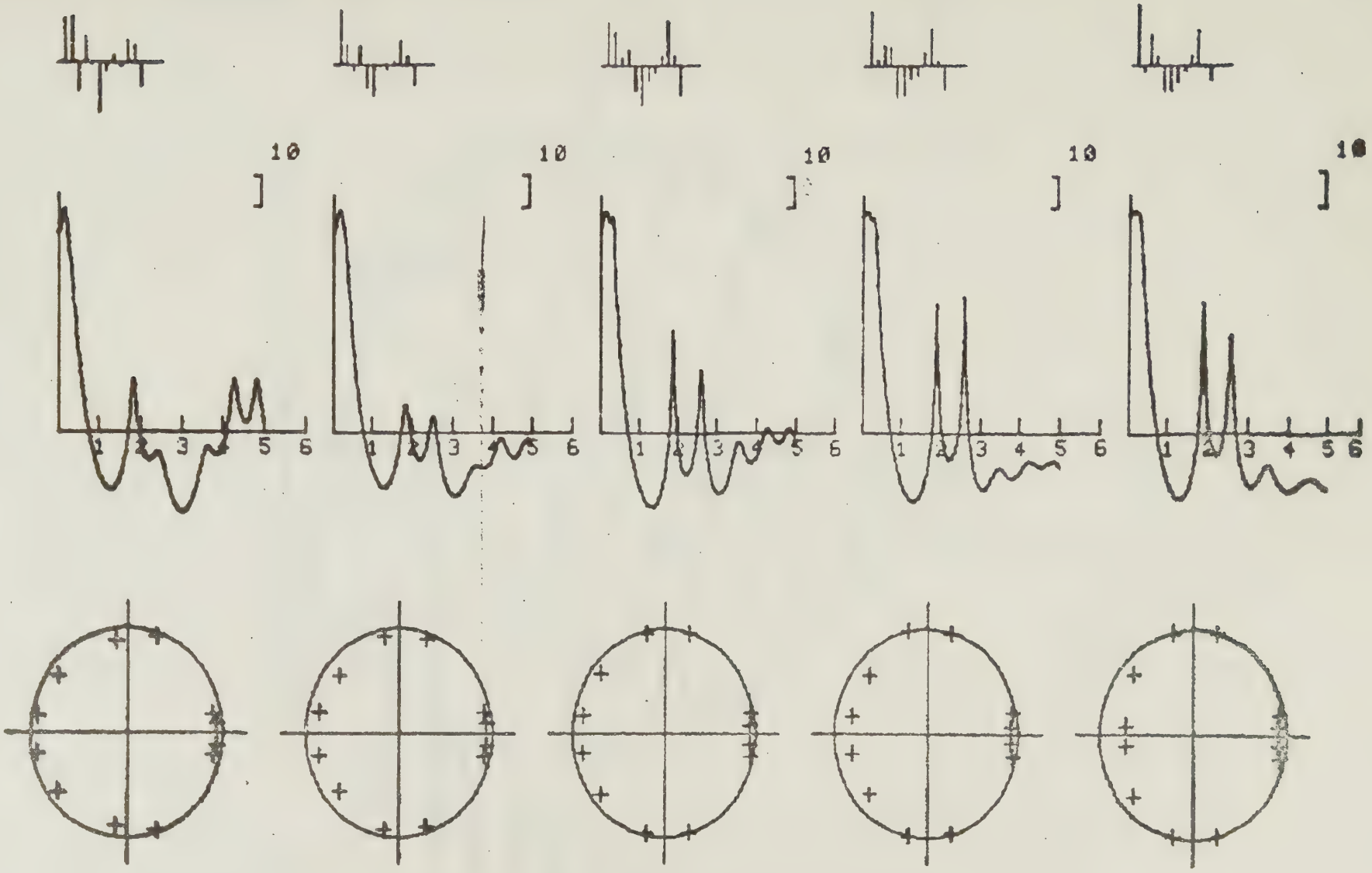


Figure A17 : Evolution des paramètres a_i , du spectre et des pôles sur la transition (/G/→/I/) pour k=12 .(calcul effectué avec le filtre de Kalman).

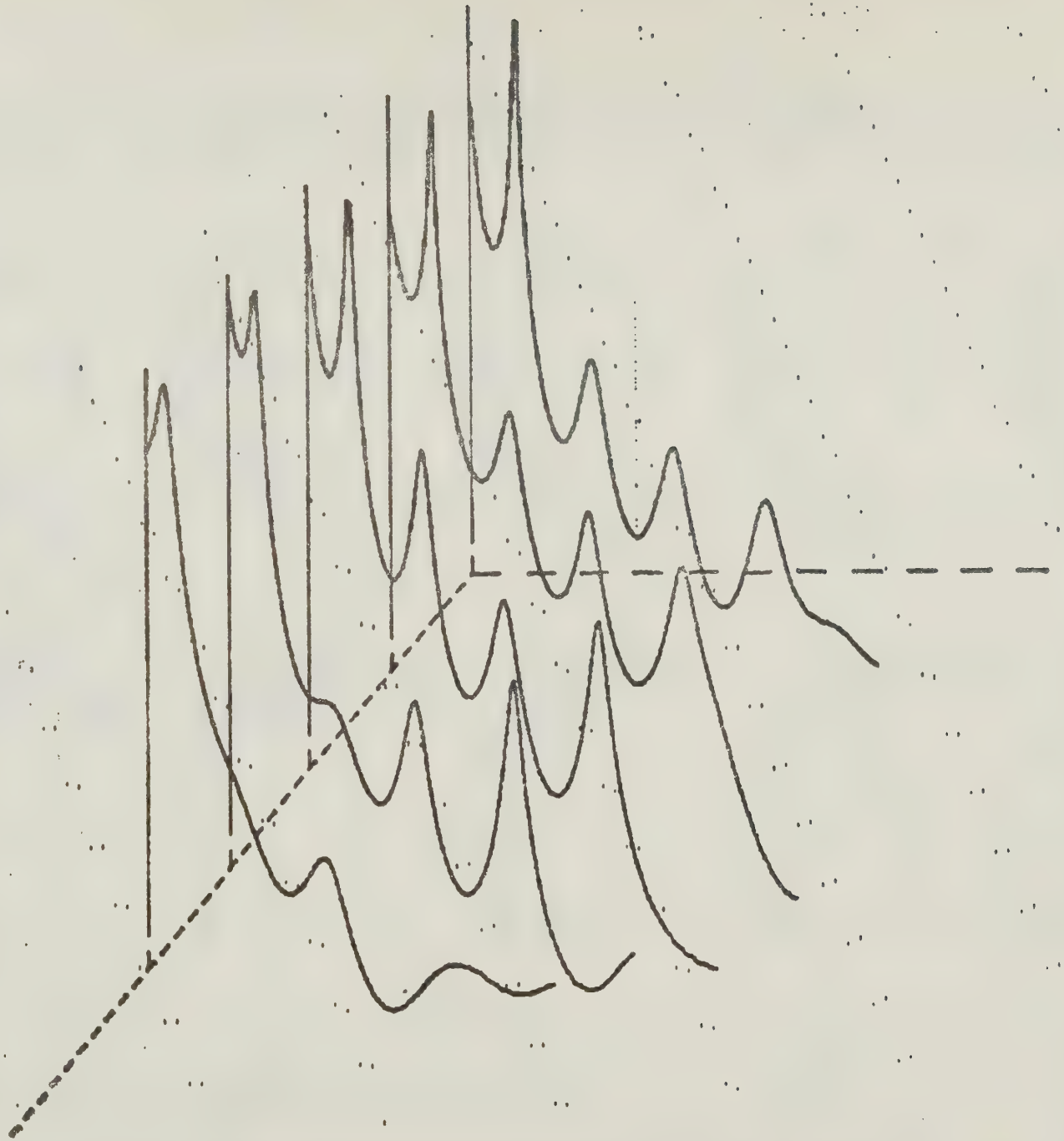


Figure A18 : Spectres evolutifs calculés avec le filtre de Kalman sur la transition /m/ →/a/ .

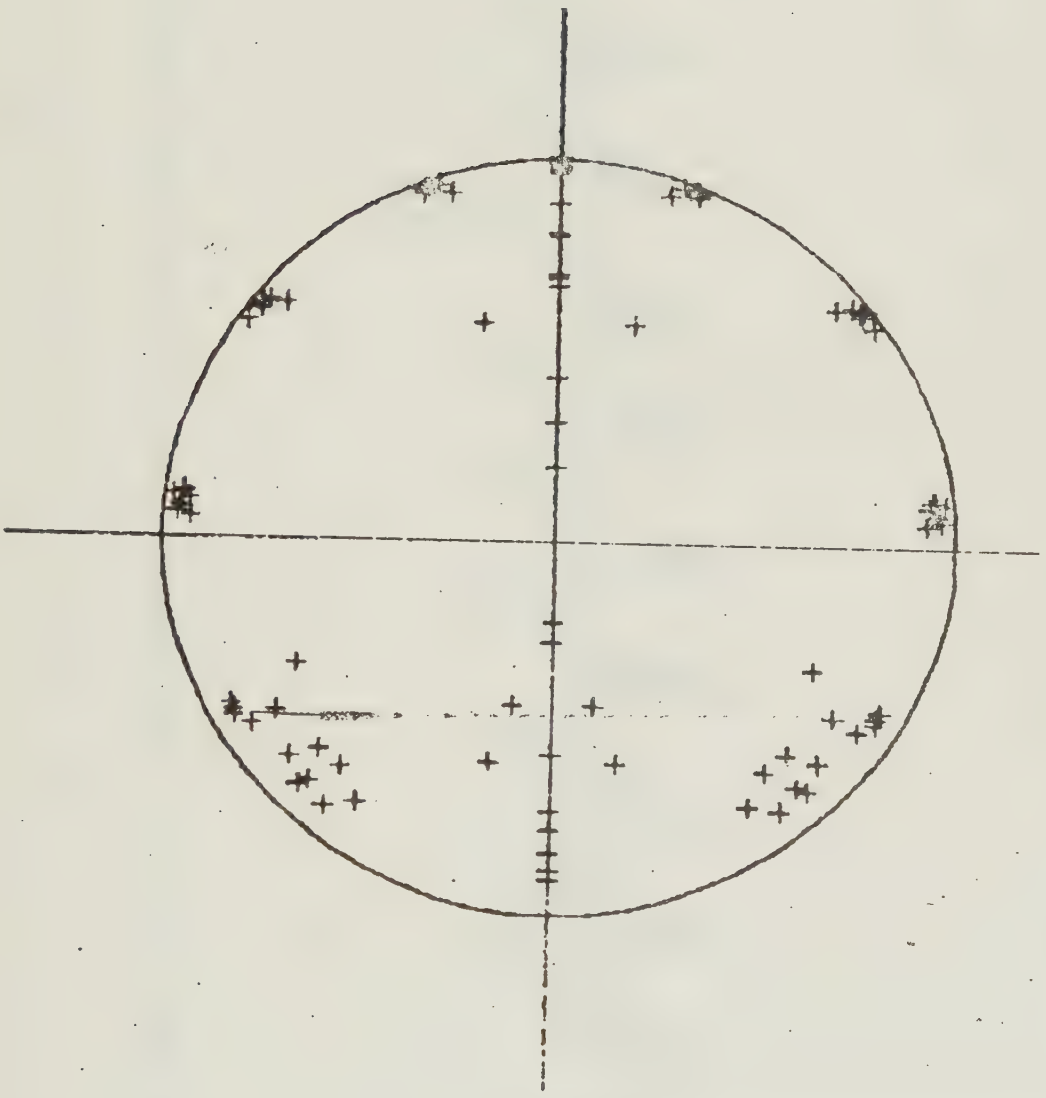


Figure A19 : Pôles du phonème /a/ calculés sur différents segments.
Les pôles les moins significatifs sont dotés d'une grande variabilité.

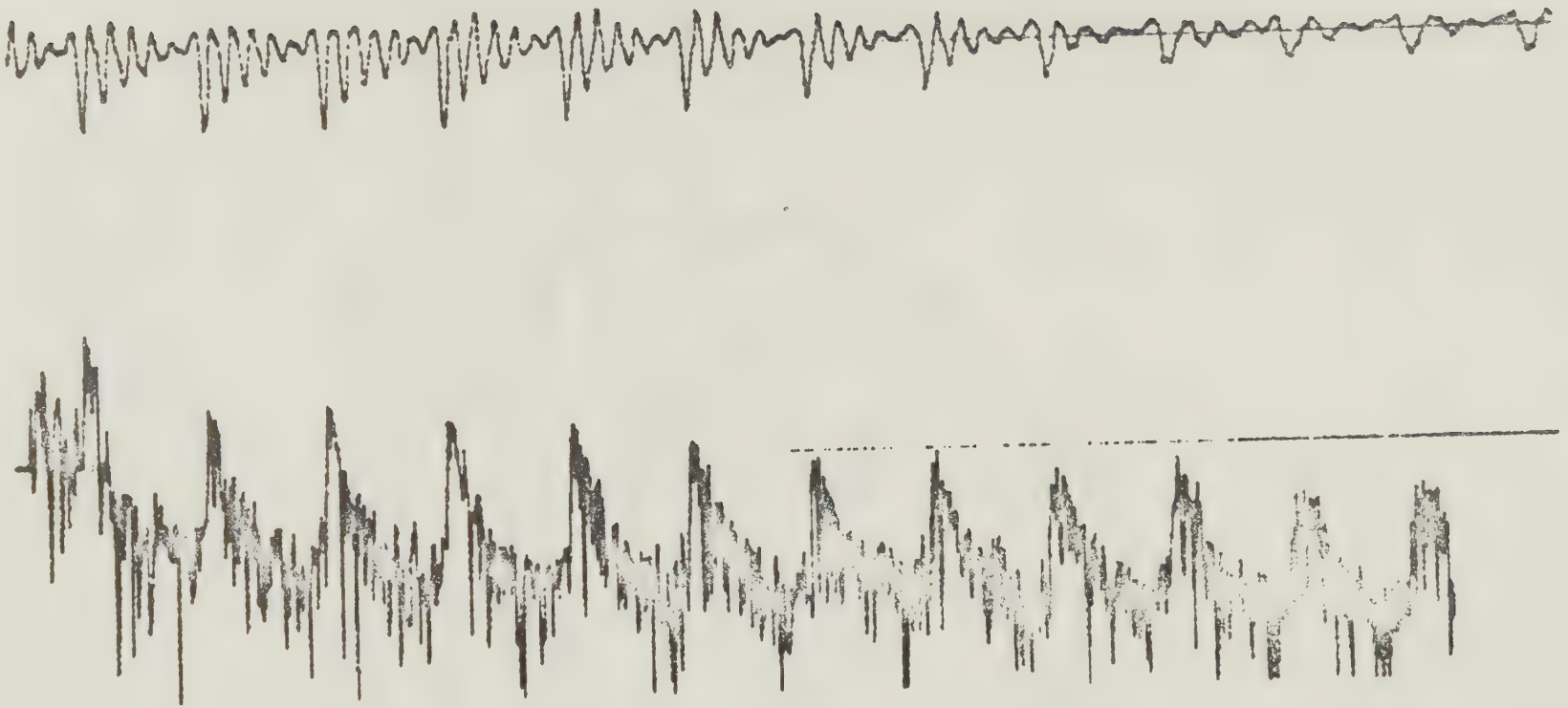


Figure A20 : Evolution du logarithme du critère γ_n sur un segment sonore d'une phrase

A N N E X E III

PROGRAMMES

(Uniquement des programmes d'analyse récurrente sont donnés dans cet annexe)

SOUS PROGRAMME D'ITERATIONS. UNE ITERATION EST EFFECTUEE A CHAQUE APPEL DE CE PROGRAMME. (MONODIMENSIONNEL).

```
      SUBROUTINE KALMAN(P,SS,A,RKK,RT,VAR,SE,SV,M)
      DIMENSION P(1),A(1),RKK(1),SS(1),RT(1)
      ID(I,J)=I+J*(J-1)/2
      MN=M-1
      RKK(1)=0.
      DØ 37 I=1,M
      L=ID(1,I)
   37 RKK(1)=RKK(1)+P(L)*SS(M-I+1)
      DØ 44 J=2,M
      RKK(J)=0.
      JJ=J-1
      DØ 45 I=1,JJ
      L=ID(I,J)
   45 RKK(J)=RKK(J)+P(L)*SS(M-I+1)
      DØ 47 I=J,M
      L=ID(J,I)
   47 RKK(J)=RKK(J)+P(L)*SS(M-I+1)
   44 CØNTINUE
      RC=0.
      DØ 46 J=1,M
   46 RC=RC+RKK(J)*SS(M-J+1)
      RC=VAR+RC
      DØ 61 J=1,M
      RT(J)=RKK(J)/RC
      DØ 61 I=1,J
      L=ID(I,J)
      P(L)=P(L)-RT(J)*RKK(I)
   61 CØNTINUE
      SE=0.
      DØ 51 J=1,M
   51 SE=SE+A(J)*SS(M-J+1)
      AASK=SV-SE
      DØ 52 J=1,M
   52 A(J)=A(J)+AASK*RT(J)
      DØ 53 J=1,MN
   53 SS(J)=SS(J+1)
      SS(M)=SV
      RETURN
      END
```

SOUS PROGRAMME D'INITIALISATION

```
      SUBROUTINE  NEØTIAL(P,S,TH,M,IND)
      DIMENSIØN P(1)
      ML=M*(M+1)/2
      IF(IND.GT.0) GØ TØ 30
      DØ 10 J=1,ML
10    P(J)=0.
      DØ 11 I=1,M
      L=I*(I+1)/2
11    P(L)=S**2
      GØ TØ 114
30    DØ 12 I=1,M
      DØ 13 J=1,I
      AA=(((J-I)/TH)**2)/2.
      L=J+I*(I-1)/2
13    P(L)=(S**2)*EXP(-AA)
12    CØNTINUE
      RETURN
      END
```

SOUS PROGRAMME D'ITERATIONS MULTIDIMENSIONNEL

```
      SUBROUTINE KALAD(A,ALPHA,MI,MU,P,SIGMA,OMEGA,IER)
      DIMENSION A(12,4),ALPHA(12),OMEGA(4),P(4,4)
      DIMENSION RMI(12,4),RKSI(200),RK(4,12),DELTAA(12)
      ID(I,J)=I+J*(J-1)/2
      MUU=MU-1
      DO 20 J=1,MI
      DO 20 I=1,MU
      RMI(J,I)=0.
      DO 20 K=1,MU
   20 RMI(J,I)=RMI(J,I)+A(J,K)*P(K,I)
      DO 30 J=1,MI
      DO 30 I=1,J
      L=ID(I,J)
      RKSI(L)=0.
      DO 30 K=1,MU
   30 RKSI(L)=RKSI(L)+A(I,K)*RMI(J,K)
      LAND=0
      DO 40 J=1,MI
      LAND=LAND+J
   40 RKSI(LAND)=RKSI(LAND)+SIGMA
      CALL SINV(RKSI,MI,EPS,IER)
      IF(IER.EQ.-1)GOTO114
      DO37I=1,MU
      RK(I,1)=0.
      DO 37 K=1,MI
      L=ID(I,K)
   37 RK(I,1)=RK(I,1)+RKSI(L)*RMI(K,I)
      DO 50 J=2,MI
      DO 50 I=1,MU
      RK(I,J)=0.
      JJ=J-1
      DO 45 K=1,JJ
      L=ID(K,J)
   45 RK(I,J)=RK(I,J)+RKSI(L)*RMI(K,I)
      DO 47 K=J,MI
      L=ID(J,K)
   47 RK(I,J)=RK(I,J)+RKSI(L)*RMI(K,I)
   50 CONTINUE
      DO 60 J=1,MU
      DO 60 I=1,MU
      DO 60 K=1,MI
   60 P(I,J)=P(I,J)-RK(I,K)*RMI(K,J)
      DO 70 J=1,MI
      OMEG=0.
      DO 71 I=1,MU
   71 OMEG=OMEG+OMEGA(I)*A(J,I)
      DELTAA(J)=ALPHA(J)-OMEG
   70 CONTINUE
      DO 80 J=1,MU
      DEL=0.
      DO 81 I=1,MI
   81 DEL=DEL+RK(J,I)*DELTAA(I)
   80 OMEGA(J)=OMEGA(J)+DEL
      DO 89 J=1,MUU
      DO 90 L=1,MI
   90 A(L,J)=A(L,J+1)
   89 CONTINUE
      DO 91 LL=1,MI
   91 A(LL,MU)=ALPHA(LL)
      RETURN
      END
```

PROGRAMME AUTOMATIQUE D'ANALYSE ET DE LA SYNTHESE DE LA PAROLE

```
      COMMON/FF/IDCB1
      COMMON/FFF/IDCB5
      COMMON/BLOCS/NW
C     DATA PII/6.283185/
      DIMENSION X(4000),IX(4000)
      DIMENSION Y(4000),IY(4000),P(220),A(20)
      DIMENSION SS(20),SP(20),SL(120),COEF(5,5),SI(400)
      DIMENSION B(20),SSF(20),SSS(20),AER(7)
      EQUIVALENCE(X(1),IX(1))
      EQUIVALENCE(Y(1),IY(1))
C
C     LECTURE SIGNAL VOCAL
C     FILTRAGE
C     CALCUL DE LA PUISSANCE TOTALE DE LA PHRASE
C
      ISEG=1
      KK=1
      IT=1
      BRU=0.
      T=0.110
      LORI=0
      READ(105,1)IDCB1,IDCB2,IDCB3,IDCB4,IDCB5,IDOW,IFOR
    1 FORMAT(7I3)
      READ(105,2)LOC,IOCC
    2 FORMAT(5I2)
      READ(105,6)IFE
      FREQC=IFE
      CALL GRIEG(COEF,FREQC,2,5,0.000055)
      CALL DAPO(LOC,IOCC,IX,IT)
      NPMAX=NW*2000
      NPM=NPMAX/2
      NWV=NW/2
      WRITE(IDCB5)LOC,IOCC,NPM,NWV
      READ(IDCB4,6)NORM
      DO 30 I=1,NWV
      CALL LECTURE(IX,IT)
      CALL LECTURE(IX(2001),IT)
      DO 46 LW=1,4000
      X(LW)=IX(LW)
   46 CONTINUE
      DO 31 J=1,4000
      K=64
      IF(I.EQ.1)K=J
      CALL FILTRE(2,5,X(J),Y(J),K,COEF)
   31 CONTINUE
      CALL ECHAN(Y,KK,2000,NORM)
      WRITE(IDCB5)(Y(J),J=1,2000)
      DO 732 J=1,2000
  732  BRU=BRU+Y(J)**2
   30 CONTINUE
      BRU=BRU/NPM
      REWIND(IDCB5)
      OUTPUT(108)'PUISSANCE TOTALE DE LA PHRASE'
      OUTPUT(108)BRU
C
C     DONNEES
C      INITIALISATIONS
```

```
      C
            READ(IDCB4,5)M
            READ(IDCB4,6)IDEB
            READ(IDCB4,6)IECHAM
            IF(IECHAM.LE.0)IECHAM=NPMAX*2000
    5       FORMAT(6I2)
    6       FORMAT(2I5,3I2)
            READ(105,2)IPMS
            PM=IPMS
            ITAN=0
            IRET=0
            IBOSS=0
            ISOU=0
            IMAX=2
            ICRIT=0
            ERDE=0.
            LEE=1
            ITTT=0
            IAK=1
            ITITI=1
            IFIN=IDEB+1800-1
            CALL LEC2(X(201),IDEB,IFIN,2000,NOE,LORI)
            READ(IDCB4,32)KK,IND,IOPT,IDECI,IMETH,ICH
            IF(IOPT.EQ.0)NN=1
            IF(IOPT.GT.0)NN=M
            MN=M-1
            ISOU=0
   32       FORMAT(I1,I1,I1,I1,I1,I1,I1)
            READ(IDCB4,66)IPP,ITH,IAA
   66       FORMAT(I6,I3,I2)
            READ(IDCB4,2)IMULT
            PW=(IPP*BRU)/0.8
            AA=IAA
            TH=ITH
            ML=M*(M+1)/2
            DO 10 J=1,M
            B(J)=0.
   10       A(J)=0.
            CALL NEGTIAL(P,PW,TH,M,IND)
            READ(IDCB4,34)VAR
            VAR=(VAR*BRU)/0.8
   34       FORMAT(F6.3)
            READ(IDCB4,2)ISAUT
            READ(105,255)W1,W2,W3
  255       FORMAT(2F6.2,F7.2)
            READ(105,1)KL,NRR,ID,IXX
            OUTPUT(108)LOC,IOCC,FREQC
            OUTPUT(108)M,KK,IDEB,IPMS,IND,IOPT,IDECI,IMETH,ICH,IPP
            OUTPUT(108)ISAUT,NORM,W1,W2,W3,PW,KL,NRR,ID,IXX
            LOC=0
            IOCC=0
            OUTPUT(108)'ANALYSE AVEC LE FILTRE DE KALMAN'
            IF(ISAUT.LE.0)ITAP=1+200;ITIP=1
            IF(ISAUT.GT.0) ITAP=ISAUT+200;ITIP=ISAUT
            IF(ISAUT.LE.0)ITON=NN
            IF(ISAUT.GT.0)ITON=NN+ISAUT
            VOR=VAR
            IECHA=(KK+1)*(IDEB+NN+ITIP-2)
```

```
      IE1=IECHA
      MM=NN+ITAP
17    DO 11 J=1,M
      IF(IOPT.EQ.0)GOTO44
      SS(J)=X(ITAP+J-1)
      GOTO11
44    SS(J)=0.
11    CONTINUE
      IF(IOPT.EQ.0)SS(M)=X(ITAP)
      SV=X(MM)
      VARI=0.
C
C     CALCUL
C
200   IECHA=IECHA+KK+1
      ITAN=ITAN+1
      ITON=ITON+1
      TON=ITON
C     TRACEP=0.
C     DO 636 J=1,M
C     L=J*(J+1)/2
C     TRACEP=TRACEP+P(L)
C636  CONTINUE
C     OUTPUT(108)TRACEP
      CALL KALMAN(P,SS,A,SP,SL,VAR,SE,SV,M)
      VARI=VARI+X(MM+1)**2
      IF(ITAN.EQ.IPMS)VARO=0.
      VARO=VARO+X(MM+1)**2
      ITITI=ITITI+1
      IF(ITITI.GE.IXX)GOTO368
370   ER=0.
      DO 70 J=1,M
      IF(IMETH.LE.1)ER=ER+(B(J)-A(J))**2
      IF(IMETH.EQ.2)ER=ER+ABS(B(J)**2-A(J)**2)
70    B(J)=A(J)
      ER=1000.*ER
      MM=MM+1
      SV=X(MM)
      IF(MM.GE.2000)GO TO 1004
1084  IF(ITAN.LT.NRR)GOTO200
      IF(ITAN.GE.(NRR+ID))GOTO1000
      AER(IAK)=ER
      IAK=IAK+1
      GO TO 200
C
C     CONSTATATION DE LA CONVERGENCE DE L ALGORITHME
C     DECISION SOURDS SONORES
C
1000  CONTINUE
      INDD=0
      IF(IONOS.EQ.1)GOTO244
      AERMOY=0.
      DO 232 J=1,ID
232   AERMOY=AERMOY+AER(J)
      AERMOY=AERMOY/ID
      DO 233 J=1,ID
233   AER(J-1)=AER(J)
      AER(ID)=ER
```

```
      IF(AERMOY.LT.W2)IONOS=1
      IF((IONOS.NE.1).AND.(ITAN.GE.KL))GOTO1002
      GOTO200
 244  IF(ER.GT.W1)INDD=1
      IF((ABS(PM-TON).GT.0.3*PM).AND.(IMAX.EG.1))INDD=0
      IF((INDD.NE.1).AND.(ITAN.GT.KL))GOTO1002
      IF(INDD.NE.1) GO TO 200
 211  IF(IMAX.EQ.1)ITTT=ITTT+1
      IF(IMAX.EQ.1)PM=(ITTT*PM+ITON)/(ITTT+1)
      IF(ITTT.GT.10)IPMS=PM
      IMAX=1
      IOCC=LOC
      LOC=INDD
      IASSON=INDD
 54   CONTINUE
 666  IF(ISOL.EQ.0)GO TO 372
      MS=MM-5
      CALL SOURCE(IBE,MS,X,SSS,M,SER,VAS,VB,VS,4C)
 372  ISOU=1
      VAR=VOR+EXP(-ABS(1,-VARI))
      VARE=VARI/ITON
      IBE=MM-M-5
C
C     SORTIES
C     SORTIE PARAMETRES SUR IDOW
C     SORTIE MESURES SUR 108
C
      WRITE(108,7)IE1,IECHA,INDD,IASSON,IONOS,VARI,VAS,VB,VS,PM
     1,MM,ICRIT
      T1=ITON
      T2=0.7*IPMS
      IF((T1.LT.T2).AND.(ISEG.GT.1))GOTO166
 167  WRITE(IDOW)M,VS,VARI,IECHA,IE1,VB,IASSON,VAS,VAR
      WRITE(IDOW)(SSS(J),J=1,M)
      ISEG=ISEG+1
 7    FORMAT(2X,2I6,I4,2I2,5F10.4,I5,F10.4,I6,I3)
C     CALCUL DES POLES ET DES FORMANTS
C     LE FICIER IFOR A ASSIGNER AU PREALABLE
C     LES SPS TRI ET POLRT A AJOUTER
C     ENLEVER LES C POUR L UTILISATION DE CETTE OPTION
C
C     A(M+1)=1.
C     DO 449 J=1,M
C449  A(J)=-SSS(M+1-J)
C     CALL POLRT(A,B,M,SSF,SP,IER)
C     M2=M/2
C     DO 439 J=1,M
CF    A(J)=ABS(1000.*ATAN(SP(J)/SSF(J)))/(PII*T)
C     B(J)=(500.*ALOG(SP(J)**2+SSF(J)**2))/(PII*T)
C     IF(A(J).EQ.0)GOTO439
C     IF(SSF(J).GT.0)GOTO439
C     A(J)=(1000./(2.*T))-A(J)
C439  CONTINUE
C     CALL TRI(A,B,M)
C     WRITE(IFOR)(A(J),J=1,M,2),(B(J),J=1,M,2)
C
C
      IE1=IECHA
```

```
      IF(IECHA.GE.IECHAM)GOTO114
C
C     REINITIALISATIONS A EFFECTUER
C
 776  IF((IDECI.GT.2).AND.(IDECI.LT.5))GOTO932
      CALL NEOTIAL(P,Pw,TH,M,IND)
      IF(IDECI.EQ.1)GO TO 933
      IF(IDECI.EQ.3) GO TO 932
 935  DO 934 J=1,M
      B(J)=0.
 934  A(J)=SSS(J)
      IF((IDECI.EQ.2).OR.(IDECI.GE.4)) GO TO 988
 932  DO 945 J=1,ML
      P(J)=IMULT*P(J)
 945  CONTINUE
      IF(IDECI.EQ.4)GO TO 935
      IF(IDECI.EQ.3)GOTO933
 933  DO 138 J=1,M
      B(J)=0.
 138  A(J)=0.
 988  INDD=0
      ITAN=0
      VARI=0.
      IONOS=0
      IAK=1
      ITON=0
      ERDE=0.
      IF(IMAX.EQ.38)GO TO200
      IF(ISAUT.LT.0)  GO TO 200
C
C     SAUT A EFFECTUER
C
      ITON=NN+ISAUT
      MM=NN+MM+ISAUT
      IECHA=IECHA*(KK+1)*(NN+ISAUT)
      IF(MM.GT.2000)IRET=12;GOTO1004
 552  CONTINUE
      IF(IRET.EQ.12)MM=MM-1800
      IRET=0
      ITAP=MM-NN
      DO538 J=1,M
      IF(IOPT.EQ.0)GOTO535
      SS(J)=X(ITAP+J-1)
      GOTO538
 535  SS(J)=0.
 538  CONTINUE
      IF(IOPT.EQ.0)SS(M)=X(ITAP)
      SV=X(MM)
      IERI=NN+ISAUT
      DO 111 J=1,IERI
 111  VARI=VARI+X(MM-J+1)**2
      GO TO 200
1004  CONTINUE
      IF(LORI.GE.NWV)GOTO114
      DO 742 J=1,200
 742  X(200-J+1)=X(2000-J+1)
      IDEB=IFIN+1
      IFIN=IFIN+1800
```

```
      IBE=IBE+1800
      IF(IBE.LT.1)IBE=1
      CALL LEC2(X(201),IDEB,IFIN,200C,N8E,L8RI)
      IF(IRET.EQ.12)G8T8552
      ICRIT=ICRIT+1
      MM=200
      G8 T8 1084
C
C     S8NS S8URDS
C
 1002 C8NTINUE
      IASS8N=0
      IF((L8C.EQ.1).8R.(I8CC.EQ.1))IASS8N=1
      IT8N=IT8N+KL+IPMS
      VARI=VARI+VAR8
      MM=MM+KL+IPMS
      IECHA=IECHA+2*(KL+IPMS)
      D8 47 J=1,M
      IF(I8PT.EQ.0)G8T8536
      SS(J)=X(MM+M+J+1)
      G8T847
  536 SS(J)=0.
   47 C8NTINUE
      IF(I8PT.EQ.0)SS(M)=X(MM+1)
      SV=X(MM)
      IMAX=38
      I8CC=L8C
      L8C=IND0
      G8 T8 54
C
  368 ITITI=1
      D8 369 J=1,M
      SSS(J)=SSF(J)
  369 SSF(J)=A(J)
      G8 T8 370
  166 ITEN=IT8N
      BACKSPACE ID8W
      BACKSPACE ID8W
      READ(ID8W)M,VS,VARI1,IECHA1,IE1,VB,IASS8N,VAS,VAR
      READ(ID0W)(SSS(J),J=1,M)
      IASS8N=0
      VARI=VARI+VARI1
      BACKSPACE ID8W
      BACKSPACE ID8W
      8UTPUT(108)'*'
      G8T8167
  114 8UTPUT(108)'FIN D ANALYSE'
      M=-9
      WRITE(ID8W)M,VS,VARI,IECHA,IE1,VB,IND0,VAS,VAR
      REWIND ID8W
      C8NTINUE
      END LABELS
C
C     SYNTHESE
C
      8UTPUT(108)'SYNTHESE'
      WRITE(IDCB3)L8C,I8CC,N8E,N
      WRITE(108,8)
```

```
 8    FORMAT(2X,'SEGMENT',4X,'PUISSANCE 1',3X,'PUISSANCE
     1,5X,'GAIN')
77    K=0
      L=0
      NRR=1
      DO638 J=1,2000
638   X(J)=0.
      ISUI=0
      READ(105,3)IND,IASSON,IOLKOS,NIOVI
3     FORMAT(5I1)
      OUTPUT(108)IND,IASSON,IOLKOS,NIOVI
      MM=IDOW
      READ(MM)M,VS,VV,IFCHA,IE1,VB,INDD,VAS,VES
      READ(MM)(A(J),J=1,M)
      READ(MM)M,VSS,VVV,IE2
      BACKSPACE MM
80    MN=M-1
      IDIF=(IECHA-IE1+NRR+1)/(NRR+1)
      IDIFF=(IE2-IECHA+NRR+1)/(NRR+1)
      ID=IDIF+IDIFF-1
      VITAL=1.
      IF(IOLKOS.EQ.2)VITAL=SQRT(VAS)
      VIT=VITAL**2
      VEN=VITAL
      IF(IASSON.EQ.1)VEN=SQRT(VIT*(1.-VB/VAS))
      D=IDIF
      VBR=VITAL/SQRT(D)
      IF((IASSON.EQ.1).AND.(INDD.EQ.1))VBR=SQRT(VB*VIT/(\
      IF((IASSON.NE.1).AND.(INDD.EQ.1))VBR=0.
      IF((IASSON.EQ.2).AND.(INDD.NE.1))VBR=SQRT(VB*VIT/(\
     1VEN=SQRT(VIT*(1.-VB/VAS))
      CALL ENTREE(SP,VEN,IND,INDD)
      IXX=101
      DO 61 J=1,MN
61    SS(J)=0.
      SS(M)=SP(1)
      II=1
      DO 40 I=1,ID
      SE=0.
      DO 51 J=1,M
51    SE=SE+A(J)*SS(M-J+1)
      CALL BRUIT(VBR,IXX,BR)
      IF(I.LT.13)SE=SE+SP(I+1)+BR
      IF(I.GE.13)SE=SE+BR
      DO 52 J=1,MN
52    SS(J)=SS(J+1)
      SS(M)=SE
      SI(M)=SE
      IF(IOLKOS.EQ.2)X(K+II)=SE
      IF((IOLKOS.EQ.2).AND.((K+II).GE.2000))GOTO81
      II=II+1
40    CONTINUE
      IF(IOLKOS.EQ.2)K=K+II-1)GOTO 18
      U=0.
      DO 38 J=1,IDIF
      U=U+SI(J)**2
38    CONTINUE
      IF(K.EQ.0)GO TO328
```

```
406!         IF(K.NE.0)GO TO 329.
407!  444    UU=0.
408!         DO 39 J=IS,ID
409!  39     UU=UU+SI(J)**2
410!         UU=(G**2)*UU
411!         WRITE(108,7)IECHA,U,UU,G,VV
412!  7      FORMAT(2X,I5,6F14.6)
413!         KK=K+1
414!         K=K+KL
415!         GOTO18
416!  81     WRITE(IDCB3)(X(J),J=1,2000)
417!         DO 89 JH=1,2000
418!         X(JH)=5000.*X(JH)
419!  89     IX(JH)=X(JH)
420!         DO 84 JF=1,2000
421!  .      IF=JF=1
422! S       LW,1     IF
423! S       LW,2     IX,1
424! S       STH,2    IY,1
425!  84     CONTINUE
426!         CALL BUFFER OUT(IDCB2,1,IY,1000,IW)
427!  85     GO TO (85,86,86,86),IW
428!  86     L=L+1
429!         OUTPUT(108)L
430!         IF(M.LE.0)GOTO114
431!         DO 636 JJ=1,2000
432!  636    X(JJ)=0.
433!         K=0
434!         KL=1
435!         II=1
436!         IF(IOLKOS.EQ.2)GOTO40
437!  .      GOTO339
438!  18     READ(MM)M,VS,VV,IECHA,IE1,V8,INDD,VAS,VES
439!         READ(MM)(A(J),J=1,M)
440!         READ(MM)M,VSS,VVV,IE2
441!         IF(M.LE.0)GOTO81
442!         BACKSPACE MM
443!         IF(NIOVI.EQ.0)INDD=0
444!         IF(NIOVI.EQ.1)INDD=1
445!         GOTO80
446!  328    G=SQRT(VV/U)
447!         KL=1
448!         DO 338 J=1,IDIF
449!         X(K+J)=G*SI(J)
450!         KL=KL+1
451!  338    CONTINUE
452!         KN=1
453!         IS=IDIF+1
454!         DO 632 J=IS,ID
455!         SL(KN)=G*SI(J)
456!  632    KN=KN+1
457!         DO 633 J=KN,120
458!  633    SL(J)=0.
459!         GOTO444
460!  329    IF(IOLKOS.EQ.1)UU=0.
461!         IF(VV.LE.UU)G=0.
462!         IF(VV.GE.UU)G=SQRT((VV-UU)/U)
463!         KL=1
```

R 10070 ** DIRECTION DE L'ENSEIGNEMENT SUPERIEUR TECHNIQUE DES TELE
UR 10070 ** DIRECTION DE L'ENSEIGNEMENT SUPERIEUR TECHNIQUE DES TELE
EUR 10070 ** DIRECTION DE L'ENSEIGNEMENT SUPERIEUR TECHNIQUE DES TELE

```
      DO 339 J=1,IDIF
      IF(IOLKOS.EQ.1)SL(J)=0.
      X(K+KL)=SL(J)+G*SI(J)
      KL=KL+1
      IF((K+KL).GE.2000)GOTO81
339   CONTINUE
      KN=1
      IS=IDIF+1
      DO 634 J=IS,ID
      SL(KN)=G*SI(J)
634   KN=KN+1
      DO 635 J=KN,120
635   SL(J)=0.
      GOTO444
114   CALL CLOSE (IDCB2)
      CALL CLOSE(IDCB3)
      STOP
      END
```

PROGRAMME CONVERSATIONNEL D'ANALYSE RECURENTE

```
      C     DOUBLE PRECISION P(220),SS(20),A(20),RA(20),RR(20)
            DIMENSION TVS(1250)
            INTEGER F1,F2
            COMMON X(2600),P,SS,B(20),RR(1024),RI(1024),RA,BR,A
            COMMON/FFF/IDCB
            IRIS=0
            NATA=0
      99    OUTPUT(108)'DCB'
            READ(105,4)IDCB
            ERDF=0.
            ITAN=0
            XC1=400.
            YC1=400.
            RAY=100.
            WRITE(108,1)
      1     FORMAT(2X,'DEBUT')
            READ(105,2)IDEB
            IFIN=IDEB+1200-1
      2     FORMAT(2I5)
            OUTPUT(108)'FREQUENCE D ECHANTILLONAGE'
            READ(105,3)IH,ITT,IBG
            IF(IH.EQ.1)T=ITT
            IF(IH.NE.1)T=ITT/1000.
            BG=IBG
      3     FORMAT(I1,I3,I4)
            OUTPUT(108)'ORDRE DU SYSTEME'
            READ(105,4)M
      4     FORMAT(3I2)
            OUTPUT(108)'OPTION'
            READ(105,32)KK,IND,IOPT
            MN=M-1
      32    FORMAT(6I1)
            OUTPUT(108)'INITIALISATIONS'
            READ(105,6)IPP,ITH
      6     FORMAT(I6,I3,I2)
            P=IPP
            AA=IAA
            TH=ITH
            OUTPUT(108)'VARIANCE'
            READ(105,34)VAR
            OUTPUT(108)'NORMA'
      34    FORMAT(F6.3)
            READ(105,2)NORM
      15    CALL IEC2(X,IDEB,IFIN,2000,NBE,IRIS)
            IF(IOPT.EQ.0)NN=1
            IF(IOPT.NE.0)NN=M
            ITAP=1
            IECHA=IDEB+(KK+1)*(NN+ITAP-2)
      345   JMAX=(IFIN-IDEB+KK+1)/(KK+1)
            CALL ECHAN(X,KK,JMAX,NBR)
            MM=NN+ITAP
            DO 11 J=1,M
            IF(IOPT.EQ.0)GOTO44
            SS(J)=X(ITAP+J-1)
            GOTO11
      44    SS(J)=0.
```

```
57:             SS(M)=X(ITAP)
58:      11     CANTINUE
59:             SV=X(MM)
60:             DA 55 J=1,MM
61:      55     X(1200+J)=0.
62:      222    ITER=0
63:             DA10 J=1,M
64:             B(J)=0.
65:      10     A(J)=0.
66:             CALL NEATTAI(P,PW,TH,M,IND)
67:      200    IF(NATA.EQ.0) NATA =1
68:             DA 405 JMC=1,NATA
69:      404    ITER=ITER+1
70:             IECHA=IECHA+KK+1
71:             ITAN=ITAN+1
72:             CALL KALMAN(P,SS,A,RA,RR,VAR,SE,SV,M)
73:      C       CALL TEARN(SS,A,SE,SV,M,VAR,ITER)
74:             TRACE=TRACEP
75:             TRACEP=0.
76:             DA636 J=1,M
77:             L=J*(J+1)/2
78:      636    TRACEP=TRACEP+P(I)
79:             DETRA=ABS(TRACEP-TRACE)
80:             VART=VART+X(MM+1)**2
81:      370    ER=0.
82:             DA 70 J=1,M
83:             ER=ER+(B(J)-A(J))**2
84:      70     B(J)=A(J)
85:             ER=1000.*ER
86:             TER=ALOG(ER)
87:             X(1200+MM)=TER
88:             NATX=10
89:             TVS(MM)=DMAXP(TER,NATX,ITER  )
90:             ERDE=0.
91:             DA 95 J=1,M
92:      95     ERDE=ERDE+A(J)**2
93:             ERDE=ER/ERDE
94:             MM=MM+1
95:             SV=X(MM)
96:             IF(MM.GE.JMAX)GATA1004
97:      405    CANTINUE
98:      33     FARMAT(I1,I3,I1,I1)
99:      28     READ(105,33)ISUT,NATA,MYKA,IAPTA
100:            IF(ISUT.EQ.5) GATA888
101:            IF(ISUT.EQ.1)GATA320
102:            IF(ISUT.EQ.3)GATA 999
103:            IF(ISUT.EQ.7)GATA666
104:            IF(ISUT.EQ.4)GATA777
105:            IF(ISUT.EQ.0)GA  TA 200
106:            IF(ISUT.EQ.6)GATA222
107:            IF(ISUT.EQ.9) GATA 53
108:     400    WRITE(108,7)IECHA,ER,TRACEP,SE,SS(M),ITAN,ERDE
109:     7      FARMAT(1X,I4,F11.4,F10.2,E7.3,I4,F10.2,F9.3)
110:     C      WRITE(108,366)(A(J),J=1,M)
111:     C366   FARMAT(F10.6)
112:            GATA28
```

```
113:      300  RR(M+1)=1.
114:           DO449  J=1,M
115:      449  RR(J)=-A(M+1-J)
116:           CALL PAIRT(RR,RA,M,RR,RI,IER)
117:           CALL TRAPAI(RR,RI,M,XC1,YC1,RAY)
118:           CALL HAMF
119:           IF(ISUI.EQ.4)GOTO200
120:           RR(1)=1.
121:           DO 661 J=1,M
122:      661  RR(J+1)=-A(J)
123:           DO 662 J=M,510
124:      662  RR(J+2)=0.
125:           CALL VITE(RR,RI,9)
126:           DO663 J=1,256
127:           RR(J)=RR(J)**2+RI(J)**2
128:           RR(J)=1./RR(J)
129:           RR(J)=20.*ALOG10(RR(J))
130:           RI(J)=((J-1)*1000.)/(512*T)
131:      663  CONTINUE
132:           XM=RI(256)
133:           YM=0.
134:           DO 664 J=1,256
135:      664  YM=AMAX(YM,RR(J))
136:           CALL TRAMI(RI,RR,256,XM,YM,200.,150.,80,750.,450.)
137:           GOTO28
138:     1004  IDEB=IFIN+1
139:           IFIN=IFIN+1200
140:           CALL LEC2(X,IDEB,IFIN,2000,NBE,IRIS)
141:           CALL ECHAN(X,KK,JMAX,NORM)
142:           MM=0
143:           GOTO28
144:      777  IF(RAY.EQ.100.)GOTO776
145:      770  XC1=XC1+150.
146:           IF((XC1.GT.900.).AND.(YC1.LE.200.))RAY=100.
147:           IF((XC1.GT.900.).AND.(YC1.LE.200.))GOTO28
148:           IF(XC1.GT.900.)GOTO778
149:           GOTO300
150:      776  XC1=-50.
151:           YC1=600.
152:           RAY=60.
153:           GOTO770
154:      778  YC1=YC1-200.
155:           XC1=100.
156:           GOTO300
157:      320  RAY=100.
158:           XC1=635.
159:           YC1=200.
160:           GOTO300
161:      888   XF=700.
162:           IF(MYKA.EQ.1)XF=1000.
163:           IF(TAPIS.EQ.1)CALL TRIC(X,JMAX,JMAX,0.,400.,XF,MATA)
164:           MARF=MM-1
165:           CALL TRIC(X(1201),JMAX,MARF,0.,150.,XF,10)
166:           CALL TRIC(TVS,JMAX,MARF,0.,600.,XF,10)
167:           GOTO 28
168:      999  REWIND IDCB
```

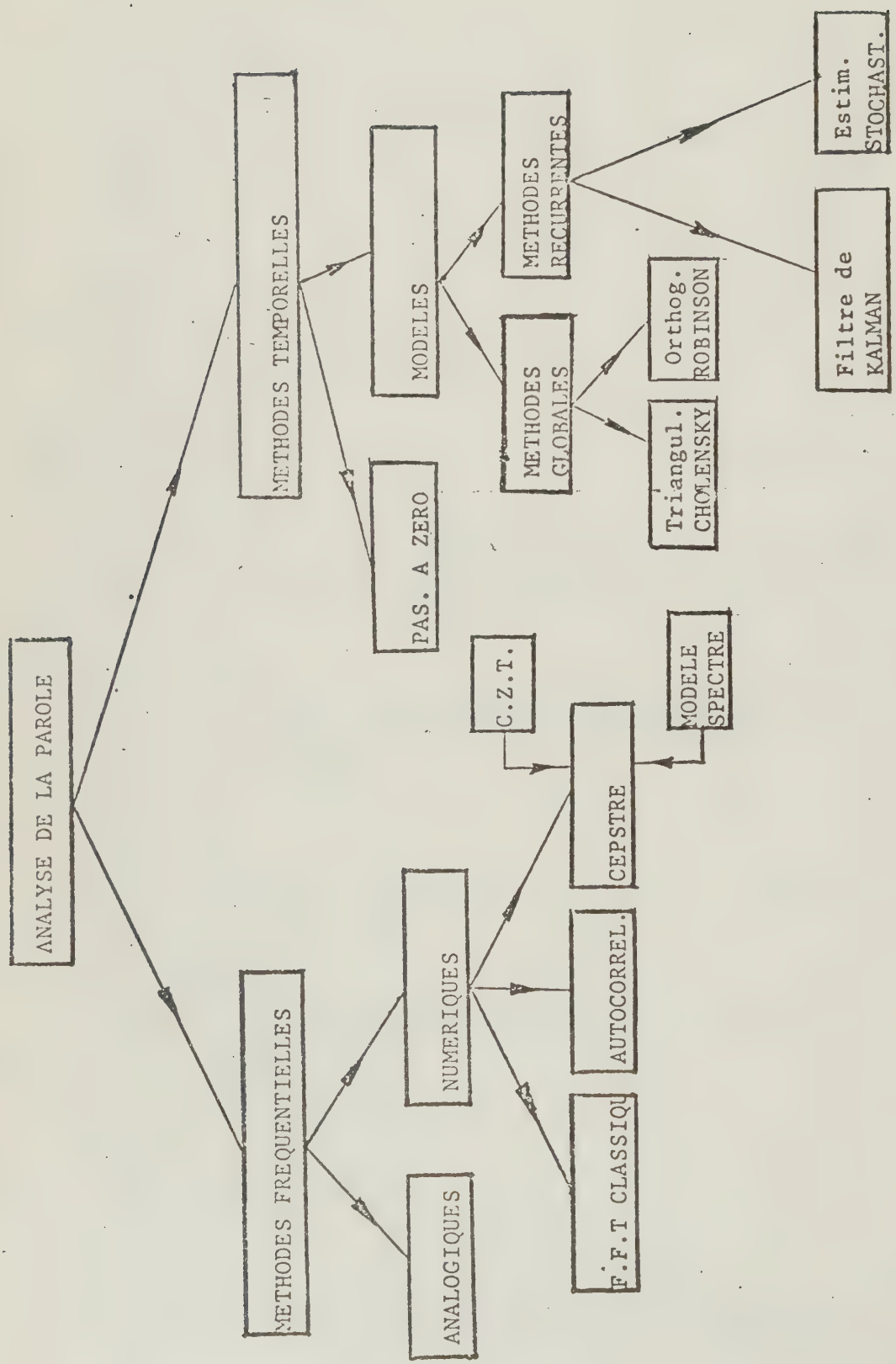
```
169:            IRIS=0
170:            GATA 99
171:      666   IF(MYKA.EQ.0) MYKA=10
172:            KF=2**MYKA
173:            KFF=KF/2
174:            E1=NATA-KFF
175:            E2=NATA+KFF
176:            IF((E1.LT.0).OR.(E2.GT.JMAX))GATA28
177:            CALL HAMMIN(RR,MYKA)
178:            DA 326 J=1,KF
179:      326   RR(J)=RR(J)*X(E1+J)
180:            CALL VITE(RR,RT,MYKA)
181:            DA327 J=1,KFF
182:            RR(J)=RR(J)**2+RT(J)**2
183:            RR(J)=20.*ALOG10(RR(J))
184:      327   RT(J)=((J-1)*1000.)/(KF*T)
185:            XM=RT(KEE)
186:            IF(TAPTA.EQ.0) GATA 330
187:             DA 329 J=1,TAPTA
188:       329   CALL MAY(RR,KFF)
189:      330   CANTINUE
190:            YM=0.
191:            DA328 J=1,KFF
192:      328   YM=AMAX(YM,RR(J))
193:            CALL TRAMT(RT,RR,KFF,XM,YM,200.,150.,BG,750.,450.)
194:            GATA28
195:       53   AUTPUT(108) 'ALPHA  (T2)'
196:            READ(105,52) TALP
197:       52   FARMAT(T2)
198:            IF(NATA.NE.0) GATA 58
199:            ALP=TALP**2
200:            DA54 I=1,M
201:            B(I)=0.
202:       54   A(I)=0.
203:       58   CALL REINT(P,ALP,M)
204:            ITER=0
205:            GATA 200
206:            END
```

```
      IDCB=0
      OUTPUT(108)'DCB'
      READ(105,1)IDCB
    1 FORMAT(4I2)
      OUTPUT(108)'SEGMENT DEBUT'
      READ(105,1)ISI
      OUTPUT(108)'ORDRES DES SYSTEMES'
      READ(105,1)MI,MU
      ITAN=MU+ISI
      OUTPUT(108)'VARIANCE'
      READ(105,3)SIGMA
    3 FORMAT(F6.3)
      OUTPUT(108)'OPTION'
      READ(105,4)IND
    4 FORMAT(3I1)
      OUTPUT(108)'INITIALISATIONS'
      READ(105,2)IPW,ITH
    2 FORMAT(I6,I3)
      PW=IPW
      TH=ITH
      DO 30 J=1,MU
      B(J)=0.
   30 OMEGA(J)=0.
      IF(ISI.EQ.0)GOTO362
      DO 363 MB=1,ISI
      READ(IDCB)
  363 READ(IDCB)
  362 DO 10 J=1,MU
      READ(IDCB)
   10 READ(IDCB)(A(L,J),L=1,MI)
      READ(IDCB)
      READ(IDCB)(ALPHA(J),J=1,MI)
  199 CALL NEOTIAL(P,PW,TH,MU,IND)
      ITER=0
  200 ITAN=ITAN+1
  201 ITER=ITER+1
      CALL KALXD(A,ALPHA,MI,MU,P,SIGMA,OMEGA,IER)
      IF(IER.EQ.-1)OUTPUT(108)'MATRICE NON DEF+'
      IF(IER.EQ.-1)GOTO114
C     OUTPUT(108)OMEGA
      ER=0.
      TRACEP=0.
      DO 20 J=1,MU
      ER=ER+(B(J)-OMEGA(J))**2
      TRACEP=TRACEP+P(J,J)
   20 B(J)=OMEGA(J)
      ER=10000.*ER
      WRITE(108,123)(ITAN,ITER,ER,TRACEP,(OMEGA(J),J=1,MU))
  123 FORMAT(2X,2I4,F12.4,F3.4,4F10.4)
C     OUTPUT(108)P
      READ(108,4)ISUI
      IF(ISUI.LE.1)GOTO367
      SOM=0.
      DO 40 J=1,MI
      SOM=SOM+ALPHA(J)
   40 CONTINUE
      SOM=5.*SOM/MI
      DO50J=1,MI
      IF(ISUI.EQ.2)ALPHA(J)=ALPHA(J)+ALPHA(J)/SOM
      IF(ISUI.NE.2)ALPHA(J)=ALPHA(J)-ALPHA(J)/SOM
   50 CONTINUE
      GOTO201
  367 READ(IDCB)
      READ(IDCB)(ALPHA(J),J=1,MI)
      IF(ISUI.EQ.0)GOTO200
      IF(ISUI.EQ.1)GOTO199
  114 STOP
```

PROGRAMME CONVERSATIONNEL
QUI GERE L'ENSEMBLE D'OPERATIONS
DU FILTRE MULTIDIMENSIONNEL

ANNEXE IV

COMPARAISON DE DIFFERENTES METHODES D'ANALYSE DE LA PAROLE


ANALYSE DE LA PAROLE
METHODES TEMPORELLES
METHODES FREQUENTIELLES
MODELES
PAS. A ZERO
ANALOGIQUES
NUMERIQUES
METHODES RECURRENTES
METHODES GLOBALES
Orthog. ROBINSON
Triangul. CHOLENSKY
Estim. STOCHAST.
Filtre de KALMAN
C.Z.T.
CEPSTRE
MODELE SPECTRE
AUTOCORREL.
F.F.T CLASSIQU

comparaison des methodes

	LOGIQUE	NOMBRE D'OPERATIONS	TEMPS REEL	TAUX DE TRANSMISSION	IMPLANTATION	ERREUR MESURE PITCH	ERREUR REPRESENTATION SYSTEME	COUT
CEPSTRE + FORMANTS	Très important	Très important	impossible	600 $\frac{bits}{sec}$	A	IO%	IO%	Très cher
CEPSTRE	Assez important	Important	Possible	7400 $\frac{bits}{sec}$	A	IO%	0	cher
CHOLENSKY	Très important	Important	Impossible	7200 $\frac{bits}{sec}$	A_S	IO%	IO%	Très cher
ROBINSON	Important	Petit	Possible	7200 $\frac{bits}{sec}$	A_S	IO%	?	Moins cher
KALMAN	Minime	Très important	Possible	7200 $\frac{bits}{sec}$	S/A	IO%	3%	Très cher

INCONVENIENTS D'ANALYSE ASYNCHRONE

I) Paramètre moyennés

2) Pitch moyenné

3) Choix d'une fenêtre (délicat)

4) Détection du pitch: impossible

AVANTAGES SUPPLEMENTAIRES DE LA METHODE BASEE SUR LE FILTRE DE KALMAN

1) Grâce au caractère recurrent : TRAITEMENT EN LIGNE avec DETECTION DU PITCH SIMULTANEE .

2) APPERCU MINUTIEUX du signal , possibilité de suivre des variations rapides (grâce à la convergence rapide de l'algorithme) . Aucune méthode ne peut donner un spectre sur un signal de durée par exemple 3 msec.

3) CONDITIONNEMENT possible de l'algorithme suivant le jeu de paramètres d'initialisation pour qu'il se prête à une tâche particulière ou pour fournir des paramètres de meilleure qualité.

INCONVENIENTS DE LA METHODE BASEE SUR LE FILTRE DE KALMAN

I) Nombre d'opérations à effectuer

très important
coût élevé

SOLUTIONS
Estimation sou optimale

2) Taux de transmission

très élevé

SOLUTIONS
Représentation avec des paramètres moins sensibles à la quantification que les a_i